Szegedi Tudományegyetem
Bölcsészettudományi Kar
Nyelvtudományi Doktori Iskola
Francia nyelvészet alprogram

Nagy Ágoston

Terminológiakivonatolás francia nyelvű szabadalmi leírásokból szabály alapú és statisztikai módszerek segítségével

PhD-értekezés

Témavezetők:

Dr. Váradi Tamás
tudományos főmunkatárs
Magyar Tudományos Akadémia
Nyelvtudományi Intézet
Nyelvtechnológiai Kutatócsoport

Dr. Gécseg Zsuzsanna
egyetemi docens
Szegedi Tudományegyetem
Bölcsészettudományi Kar
Francia Nyelvi és Irodalmi Tanszék

Szeged
2012

Tartalom

Bevezetés....1

1. A tudományterület meghatározása....6

1.1. A terminológia szó definíciói....6

1.2. A terminológia kialakulásának kezdetei....8

1.3. A terminológia mint interdiszciplináris terület....10

1.3.1. Terminológia és nyelvészet....11
1.3.2. Lexikológia és terminológia: két külön terület?....11
1.3.3. Terminológia és szabványosítás....13
1.3.4. A terminológia kapcsolata más tudományterületekkel....16

1.4. A terminológia különböző területekre történő tagozódása....17

1.5. A disszertáció területe....20

2. A szaknyelv jellemzői....21

2.1. A szaknyelv viszonya más nyelvi rétegekkel....21

2.2. A szaknyelv lexikai és szintaktikai jellemzői....23

3. A terminusok jellemzői....27

3.1. A terminus klasszikus definíciója....27

3.1.1. A fogalom definíciója....28
3.1.2. A szakterület fogalma....29
3.1.3. A terminus-fogalom közötti bijektív kapcsolat....29

3.2. A terminusok további meghatározásai....30

3.3. A terminusok összehasonlítása a köznyelvi egységekkel....32

3.4. A terminusok definíciói az automatikus terminológiakivonatolásban....33

3.5. A terminusok jellemzői....35

3.5.1. A terminusok összetettsége....35
3.5.2. A terminusok morfológiai jellemzői....37

3.6. A terminus definíciója a saját TE-alkalmazásban....38

4. A francia nyelvű főnévi terminusok szerkezete....39

4.1. A francia főnévi csoportok szerkezete....40

4.1.1. Prepozíciós szintagmából álló komplementumok....42
4.1.2. Melléknevek helye a főnévi csoportban....44

4.2. A francia főnévi terminusok belső szerkezete....49

4.2.1. Prepozíciós szintagmák a főnévi terminusokban....49
4.2.2. Melléknevek helye a főnévi terminusokban....53
4.2.3. Determinánsok és egyéb adjunktumok a terminusokban....54

4.3. A többszavas francia terminusok belső szintaktikai jellemzői....56

5. A terminológiakivonatolás módszerei....59

5.1. A terminológiakivonatolás célja....59

5.1.1. Fordítói munka elősegítése....60
5.1.2. Dokumentumok indexelése....61
5.1.3. Terminológiai adatbázisok létrehozása....61
5.2. Terminológiakivonatolás lépései....62
5.2.1. Korpusz/tudományterület kiválasztása....63
5.2.2. A korpusz szegmentálása, nyelvi elemzése....65
5.2.2.1. Korpusz szegmentálása mondatokra....65
5.2.2.2. Korpusz szegmentálása tokenekre....66
5.2.2.3. Tokenek szófaji címkézése és lemmatizálása....66
5.2.2.4. Szintaktikai elemzés....67
5.2.3. Terminusjelölt-lista összeállítása....68
5.2.4. Terminusjelölt-lista szűrése....68
5.2.5. Validálás....69
5.3. A terminológiakivonatolók szabály alapú moduljai....71
5.3.1. Mintaillesztés....71
5.3.1.1. Mintaillesztés reguláris kifejezéssel....72
5.3.1.1.1. Bevezetés a reguláris kifejezések használatába....72
5.3.1.1.2. Reguláris kifejezések használata terminológiakivonatoláshoz....73
5.3.1.2. Mintaillesztés véges állapotú automatával....74
5.3.2. Újraíró szabályok terminusvariánsok kinyerésére....77
5.3.3. Konnektívumok szűrése....78
5.3.4. Terminológiai helyzet....80
5.4. A terminológiakivonatolók statisztikai moduljai....81
5.4.1. Gyakoriság vizsgálata – termhood-mértékek....82
5.4.1.1. TF-IDF....83
5.4.1.2. Log-likelihood....85
5.4.1.3. Weirdness....86
5.4.2. Unithood-mértékek....87
5.4.2.1. Mutual Information....87
5.4.2.2. C-érték....88
5.4.2.3. Mutual Expectation....89
5.4.2.4. IR/CR....90
5.4.2.5. Egyéb mértékek....91
5.4.3. Kontextus figyelése....92
5.4.4. Tulajdonnevek szűrése....95
5.5. Terminológiakivonatolók összehasonlítása....97
5.6. Francia nyelvre készült terminológiakivonatolók....99
5.7. Saját terminológiakivonatoló megvalósítása: célok és hipotézisek....101
6. A korpusz bemutatása....104
6.1. A szabadalmak részei....104
6.1.1. Bibliográfiai adatok....104
6.1.2. Leírás és igénypontok....106
6.2. A kiválasztott korpusz....107
6.3. A korpusz kinyerése....111
6.4. A korpusz annotálása....113
7. Terminológiakivonatolás megvalósításának módszere....114
7.1. A korpusz előfeldolgozása....114

7.1.1. Korpusz mondatokra, majd tokenekre bontása, és azok címkézése....114
7.1.2. Tokenek kezelése....116
7.2. Szűrők....118
7.2.1. Tulajdonnevek szűrése....118
7.2.2. Konnektívumok és egyéb szókapcsolatok szűrése....119
7.3. Automata létrehozása mintaillesztés céljából....120
7.4. Minta illesztése....123
7.5. Statisztikai módszerek....124
7.5.1. Termhood-érték kiszámítása....125
7.5.1.1. A világhálón történő keresés megvalósítása....125
7.5.1.2. Weirdness-érték kiszámítása....127
7.5.2. Unithood-érték kiszámítása....130
7.5.3. Terminusok súlya....132
7.5.4. Összevont érték....135
7.5.4.1. A három mérték közös tartományba történő leképezése....135
7.6. Adatbázis létrehozása....138
8. Eredmények....142
8.1. Az eredmények számításának módszere....142
8.2. Szabály alapú kinyerés és szűrés eredményei....143
8.3. A terminológiakivonatolás hatékonyságának növelése statisztikai módszerekkel....144
8.3.1. Az egyes értékek súlyozása az összevont érték esetében....145
8.3.2. Statisztikai értékek hatása a terminológiakivonatolás eredményeire....146
8.3.3. A korpusz nagyságának megfelelősége....147
8.4. A terminológiakivonatolás eredményei az A23L korpuszon....148
8.5. A terminológiakivonatolás eredményeinek elemzése....151
8.5.1. A szabály alapú kinyerés és szűrés hatékonysága....151
8.5.2. A szabály alapú kinyerés és szűrés lehetséges hibaforrásai....153
8.5.2.1. Helytelen morfológiai annotációk....153
8.5.2.2. Egyéb hibaforrások....154
8.5.3. A statisztikai módszerek eredményei....155
8.6. Összevetés más terminológiakivonatoló alkalmazásokkal....156
9. Összegzés és további kutatási lehetőségek....160
Köszönetnyilvánítás....165
Források....166
Bibliográfia....167
Mellékletek....174

Bevezetés

Az automatikus terminológiakivonatolás (TE – *term extraction*) során egy erre a célra készített alkalmazás egy adott, írott nyers szakszövegből kinyeri annak terminusait. A terminológiakivonatoló alkalmazások leginkább csak más programok kiegészítéseként használatosak (ahogyan a helyesírásellenőrzők is a szövegszerkesztő programok elengedhetetlen részei), tehát önálló alkalmazásként csak ritkán, ennek ellenére napjainkban ez a számítógépes nyelvészet egyik igen kutatott területe.

A TE egyik alkalmazási célja az automatikus szövegindexelés, amely során egy megadott szöveges fájl rá jellemző fontosabb kifejezéseit kivonatoljuk, amelyek a későbbiekben elősegíthetik ezen dokumentumok kategorizálását vagy a köztük történő keresést. Erre az eszközre épül például az internetes keresőmotorok egy része is, amelyek a már indexelt honlapokat tárolják a gyorsabb és hatékonyabb keresés megvalósítása érdekében (Enguehard 2005). Ugyanakkor szintén terminológiakivonatoló eszközökre támaszkodnak gépi fordítást megvalósító alkalmazások (pl. Vasconcellos 2001), illetve információkinyerő eszközök (pl. Ahmad 2001).

A TE egy másik felhasználási területe a fordítói munka elősegítése. Ha például egy terjedelmes szakmai könyvet (például egy szoftver vagy hardver leírását) kell rövid idő alatt konzekvensen lefordítani, azt általában egy fordító nem tudja megoldani. Ilyenkor fontos lehet egy terminusjelölt-lista, amit, ha a lektor előre kézhez kap, már a szöveg fordítóknak történő kiadása előtt lehetősége lenne megadni az abban szereplő terminusok idegen megfelelőit. Így a csoportban dolgozó fordítók feltehetően minden terminust ugyanúgy fognak fordítani.

A TE, akárcsak a számítógépes nyelvészet szinte minden területén, történhet szabály alapú és statisztikai módszerekkel, azonban ez nem jelenti azt, hogy ezen alkalmazások csak az egyikre támaszkodnának: a többségük a kettőt ötvözi (Maynard és Ananiadou 2000). Cabré és mtsai (2001) szerint nem is javasolt, hogy a két módszer közül csak az egyiket alkalmazzuk, mivel a szabály alapú kivonatolók túl nagy zajt (tehát a kivonatolt terminusjelöltek száma magasabb, mint a valós terminusoké), a statisztikai alapúak pedig túl nagy csendet okoznak (a terminusjelöltek listája sok terminust nem tartalmaz). A szabály alapú módszer esetében a terminusok a belső morfoszintaktikai szerkezetük segítségével kivonatolhatók, tehát terminusokra jellemző összetételeket kell keresnünk, például két egymás után álló főnevet. A statisztikai kivonatolók olyan lexikai

egységeket keresnek fel a szövegekben, amelyek gyakrabban fordulnak elő együtt a többi egységhez képest. Ezt még ki is lehet egészíteni azzal, hogy ezután minden nem üres szópárra megmérhetjük annak gyakoriságát egy hétköznapi nyelvet tükröző referenciakorpuszban. Ha a gyakoriság a szaknyelv javára nagyon eltér, akkor igen nagy valószínűséggel terminusról van szó. A statisztikai alkalmazások gyakran használnak asszociációs mértéket is, ugyanis azok az igazi terminusi összetételek, amelyek egy szövegben egymáshoz mindig közelebb vannak, illetve gyakrabban fordulnak elő együtt, mint külön. Ezen kívül sok esetben árulkodó a terminusok szövegkörnyezete, mivel vannak olyan szerkezetek, amelyek nagy valószínűséggel követnek vagy előznek meg terminusokat (pl. *la notion de* … 'a … fogalma').

Cabré és mtsai (2001), valamint Sauron (2002) szerint a TE legfőbb szakaszai a terminusok kinyerése, majd azok szűrése. Azaz bármelyik módszert választjuk, az első szakaszban kinyert terminusok között biztos vannak olyan jelöltek, amelyek nem terminusok, és amelyeket ezáltal szűrni kell. A terminusjelöltek kinyerése és szűrése leggyakrabban szabály alapú és statisztikai módszerek kombinációjával történik (hibrid módszerrel): Cabré és mtsai (2001), valamint Ha és mtsai (2008) szerint először a terminusjelöltek kinyerésére a statisztikát alkalmazzuk, majd azok szűrésére nyelvi filtereket.

A disszertáció célja, hogy bemutassa egy általunk létrehozott terminológiakivonatoló működését, eredményeit. Az alkalmazás, amely tartalmaz mind statisztikai, mind szabály alapú modulokat, – az utóbbiak miatt – csak francia nyelvű szövegeken futtatható. A terminológiakivonatolás korpuszának francia nyelvű szabadalmi leírásokat választottunk, így a terminuskinyerési szabályokat és a statisztikai módszereket erre optimizáltuk.

Vizsgálatunk során nem a szokásos (statisztikai alapon történő terminuskinyerés, majd szabály alapú szűrés) eljárást alkalmaztuk, hanem ennek fordítottját, ami ritkább a TE során: a terminusjelölt-listát szabály alapú módszerekkel nyertük ki, majd ezt különböző szűrőkkel szűrtük. A terminusjelöltek szabály alapú kinyerésének választásával azt kívántuk bizonyítani, hogy az a szakirodalomban elterjedt nézet (pl. Cabré és mtsai 2001), miszerint a statisztikai módszerekkel történő terminuskinyerés hatékonyabb, nem feltétlenül igaz minden esetben. E választást az is igazolja, hogy (1) a francia nyelvben a főnévi terminusok nagy arányban rendelkeznek olyan belső szerkezettel, ami a köznyelvi egységekre nem jellemző, így a szabály alapú módszerek nagyobb eredményességgel

használhatók; valamint az, hogy (2) a francia nyelvre is alkalmazható, legtöbbet idézett terminológiakivonatolókban a kinyerés szabály alapú (5.6. fejezet).

A szöveget a szabály alapú kinyerés előtt szintén szabály alapon különböző *stopword*ökkel szűrtük, azaz a szövegben előforduló szavak közül kiszűrtük azokat, amelyek olyan elemeket tartalmaznak, amelyek nem lehetnek terminusok részei. A szabály alapon kinyert terminusjelölt-listát végül statisztikai módszerekkel tovább szűkítettük. Célunk ezzel az volt, hogy megmutassuk, hogy a különböző szabály alapú, illetve statisztikai szűrők milyen hatékonysággal járulnak hozzá a szabály alapú terminuskinyerés által létrehozott terminusjelölt-lista megfelelőségének javításán. Előfeltételezéseink szerint a szabály alapú TE a leírásokban szereplő terminusok nagy részét kinyeri, de a listába valószínűleg sok olyan terminusjelölt kerül, amely nem terminus. A szabály alapú terminuskinyerés által okozott zajt pedig mindkét típusú szűrővel jelentősen lehet javítani anélkül, hogy a helyesen kinyert terminusok száma lehetőleg ne vagy csak alig csökkenjen. A disszertáció egyik célja az volt, hogy megmutassuk, milyen mértékben járulnak hozzá a különböző faktorokat figyelembe vevő statisztikai mértékek a szabály alapú kinyerés hatékonyságának növeléséhez, valamint kidolgozzunk egy olyan módszert, amely az alkalmazott három statisztikai mértékből egy összevont értéket képez. Célunk volt még megállapítani, hogy az összevont statisztikai értéknek milyen küszöbértéket kell adnunk ahhoz, hogy a pontosságot jelentősen meg tudjuk növelni a fedés lehető legkisebb csökkenésével.

A saját terminológiakivonatoló olyan szabály alapú (terminusok belső szintaktikai összetételét vizsgáló) és statisztikai (*weirdness,* C-érték, súly) modulokat használ, amelyeket más terminológiakivonatolók is alkalmaznak, de nem feltétlenül ebben a kombinációban vagy arányban (pl. Frantzi és Ananiadou (1999) szintén kombinálja a C-értéket a súlyértékkel, de csak ezt a kettőt alkalmazza a statisztikai modulok közül és nem a jelen disszertációban leírt arányban), illetve nem ebben a sorrendben (a kinyerés szabály, a szűrés statisztikai alapú). A disszertáció egyik jelentősége az, hogy az adott korpuszra megmutatja, milyen hatékonysággal járul hozzá minden egyes használt eljárás (mind a szabály alapú mind a statisztikai) külön-külön és együttesen a TE hatékonyságának növeléséhez.

A terminológiakivonatoló alkalmazás eredményeit ezenkívül más – hasonló feladatot (pl. szövegindexelés) elvégző – programok eredményeivel (Fastr és YaTeA) is összevetettük. Annak ellenére, hogy ezen alkalmazások a kinyerést szabály alapon

valósítják meg, azért ezen programokat használtuk, mert szabadon elérhetők. Statisztikai alapú kinyerést megvalósító alkalmazások csak korlátozottan érhetők el (pl. csak az első húsz terminust mutatják meg) és/vagy csak statisztikai értéket rendelnek a terminusjelöltekhez, de a küszöbértéket, amely felett egy terminusjelölt lehetséges terminus, azt nem (pl. a Translated.net Labs *Terminology Extraction* programja[1]). A disszertációban a 158. oldalon szereplő 8.8. táblázat összeveti az általunk kidolgozott eljárás hatékonyságát a YaTeA és a Fastr rendszerekével: a YaTeA és a Fastr átlagos pontossága 0,36, fedése 0,57, a saját terminológiakivonatoló ezen értékei 0,62 és 0,79. Ez egyrészről azt mutatja, hogy számít a választott szófaji címkéző program (a két külső terminológiakivonatoló által alkalmazott szófaji címkéző program nem ugyanaz, mint a saját alkalmazásban használt), másrészről a francia nyelvű szabadalmi korpuszunkban a szabály alapú kinyerés hatékony.

A disszertáció első öt fejezete adja meg az empirikus kutatás elméleti hátterét. Az első fejezet célja a terminológiával kapcsolatos tudományterületek rövid ismertetése és az automatikus terminológiakivonatolás területének behatárolása. A második fejezetben egy konkrét példa segítségével mutatjuk be a szaknyelv főbb jellemzőit, amelynek célja az, hogy ezen kritériumok alapján válasszunk korpuszt. A harmadik fejezetben ismertetjük a terminusok különböző definícióit és jellemzőit. Erre a fejezetre azért van szükség, hogy felvázoljuk azokat a kritériumokat, amelyek alapján a saját terminológiakivonatolót létrehoztuk, illetve ezek azok a feltételek, amelyek alapján eldönthető, hogy egy lehetséges kinyert terminusjelölt az-e vagy sem.

A negyedik fejezetben ismertetjük a francia nyelvű főnévi csoportok szerkezetét, mivel a terminusok egyik fő ismertetőjegye, hogy nem köznyelvi egységek, így a hagyományos köznyelvi főnévi csoportok főnévi terminusokkal történő összevetése támpontot adhat a szabály alapú terminuskinyeréshez szükséges főnévicsoport-minták kidolgozásához, amelynek során cél, hogy a terminuskinyerés során minél kevesebb nemterminusi elem kerüljön a listába. Az ötödik fejezetben egy általános kitekintést nyújtunk a korábbi, illetve aktuális terminuskinyerési módszerekről. A fejezet első részében írjuk le ennek a feladatnak a lépéseit a korpusz előfeldolgozásától kezdve a validálási folyamatig. Ezt követően a nemzetközi szakirodalom alapján részletesen bemutatjuk a terminológiakivonatolásban használt szabály alapú és statisztikai módszereket.

[1] http://labs.translated.net/terminology-extraction

A hatodik fejezetben ismertetjük az elemzéshez választott korpuszt, amely kizárólag francia nyelvű szabadalmak leírásából áll. Ahhoz, hogy a TE ne csak egy szűk szakterületre korlátozódjék, kétféle szabadalmat választottunk: az egyik a G06F-osztály, ami a digitális adat feldolgozásának területébe tartozó szabadalmakat tartalmazza, a másik az A23L-osztály, amely élelmiszerekkel, élelmiszerfélékkel vagy alkoholt nem tartalmazó italokkal foglalkozik. Mindkét szabadalmi területről 10-10 leírást választottunk, amelyek átlagosan 4000 szövegtokent tartalmaznak. Azért választottuk a szabadalmak ezen részét, mert ezekben sok ismétlés található, így a statisztikai módszerek is hatékonyabbak lehetnek.

A hetedik fejezetben mutatjuk be az általunk kidolgozott terminológiakivonatoló alkalmazás működését, a használt szabály alapú, illetve statisztikai módszereket. Az előbbi kategóriába tartozik a véges állapotú, determinisztikus automatával történő szabály alapú terminuskinyerés és a *stopword*ök szűrése, az utóbbi kategóriához a szövegkörnyezetet figyelembe vevő súlyérték (Frantzi és Ananiadou 1999), a terminusjelöltek köznyelvi korpuszhoz mért gyakorisága, azaz a *weirdness*-érték (Ahmad és mtsai 1999), valamint a terminusjelöltek elemeinek összetartozását mérő C-érték (Frantzi és Ananiadou 1999).

Ezen fejezetet az eredmények leírása követi (8. fejezet), amelyben először azt mutatjuk meg, hogy milyen hatékonysággal történt a tisztán szabály alapú kinyerés, majd azt, hogy ezen milyen mértékben tudtak változtatni a szabály alapú és statisztikai szűrők. A dolgozatot az összegzés és a további lehetséges feladatokat bemutató rész zárja.

A disszertáció és a disszertációhoz tartozó program – az importált modulokon kívül – saját, egyedül végzett munka eredménye. A terminológiakivonatoló alkalmazást objektumorientált Java programozási nyelven írtuk Eclipse fejlesztői környezetben: a disszertációhoz mellékelt CD tartalmazza a program forráskódját, illetve annak fordított, gépfüggetlen bájtkódját, valamint a mintakorpuszt, illetve az eredményeket tartalmazó fájlokat. A program mind Windows XP, mind Linux operációs rendszer alatt is működik.

1. A tudományterület meghatározása

E fejezet célja a terminológiával kapcsolatos tudományterületek rövid ismertetése és a disszertáció területének behatárolása. Ahogy azt ebben a részben bemutatjuk, a TE több tudományterület együttes ismeretét és használatát igényli, ezen belül is a nyelvészetét, a matematikai statisztikáét és a mesterséges intelligenciáét. Ez már mutatja a terület összetettségét, amely miatt a terminológiakivonatolási tevékenység egy interdiszciplináris megközelítést igényel.

Először a terminológia fogalmát vezetjük be (1.1.), majd röviden ismertetjük a terminológia tudományának rövid történetét (1.2.). Az ezt követő 1.3. fejezetben bemutatjuk, milyen sok területből áll össze a terminológia, és azt, hogyan vezetnek a különböző terminológiai problémák a terület felaprózódásához (1.4.). Végül (1.5.) azt ismertetjük, hogy milyen fő ágai vannak a modern terminológiának, és hogy ezekhez hogyan kapcsolódik a TE.

1.1. A terminológia *szó definíciói*

A *Magyar nagylexikon* (Vizi 2003, 17. kötet, 367) szerint a *terminológia* szó egyjelentésű, jelentése „szakszókészlet: valamely szakmában v. tudományágban általánosan használt és a köznyelvben is előforduló szakszók összessége”, tehát minden egyes tudományterületnek, például az orvostudománynak, az informatikának, a biotechnológiának létezik saját terminológiája. Az első definíció így a terminológia fogalmát egy tudományterület *terminus technicus*ainak egy halmazára redukálja.

A Kovács (2001: 22) által adott definíció csak az egyik jelentést mutatja be, így az azonos nevet viselő tudományterület nem kerül be a *terminológia* szó meghatározásai közé:

> Az egyes tudományágak terminológiája mindazon szavak és kifejezések, szerkezetek összessége, amelyek az adott tudományág legáltalánosabb jellemző vonásait, folyamatait megközelítő pontossággal tükröző fogalmak megnevezésére szolgálnak, viszonylag zárt rendszert alkotnak, és adott nyelvi helyzetben környezetfelidéző hatással rendelkeznek.

Itt érdemes kitérni a meghatározás több – pontatlanságot sugalló – kifejezésének részletes magyarázatára is, amelyek közül az első szembetűnő a „megközelítő pontossággal tükröző fogalmak” szerkezet. A fenti meghatározásból úgy tűnik, hogy az adott nyelvi jelek kidolgozása nem lenne eléggé precíz, de valójában arról az alaptételről van szó, hogy az elsajátítás előrehaladtával a fogalmak „egyre pontosabban, de sohasem tökéletes

pontossággal tükrözik az adott jelenség, folyamat jellemző vonásait az emberi megismerő agyban." (Kovács 2001: 22–23).

A második részletezendő kifejezés a „viszonylag zárt rendszer". Mint ahogy azzal a 2. fejezetben is foglalkozunk, nagyon nehéz élesen elhatárolni egymástól a köznyelvet és a szaknyelvet (Lérat 1989), hiszen sok hétköznapi kifejezés válik terminussá (pl. az informatikában használt *háló*, *szemétgyűjtés* szavak), illetve sok terminus válik egy idő után hétköznapi kifejezéssé (pl. a műszaki területekből a technikai fejlődés során átkerült kifejezések, pl. *gramofon, plazmatévé, MMS*). Így ez azt jelenti, hogy ugyanazon szó vagy mást jelent, vagy egyben már nem is terminus, ha köznyelvi szövegekben alkalmazzuk.

A „környezetfelidéző hatás" (*effet par évocation du milieu*) a terminusok létrehozásának arra az elvére épül, miszerint a terminusok általában köznyelvi szavakból épülnek fel, ezért azok minél inkább megpróbálják az adott fogalmat, jelenséget a lehető legpontosabban leírni. Az informatikából említett *világháló* fogalma jól tükrözi az internet szerteágazóságát, és a *szemétgyűjtés* a felszabadított memória képzetét a megtisztított parkokéhoz vagy járdákéhoz társítja.

Más egynyelvű lexikonokban (pl. *Magyar értelmező kéziszótár*, Csábi 2007, 1335) azt tapasztaljuk, hogy a terminológia fogalmába nem építik be az ezzel kapcsolatos tudományterületet, amelyet szintén ugyanígy nevezünk. Fóris (2005) ezért idegen nyelvű cikkre hivatkozik a terminológia fogalmának meghatározásakor, és elismeri, hogy ebbe a szócikkbe beletartozik a tudományterület is. A leghíresebb és legtöbbet hivatkozott francia egynyelvű kéziszótár, a *Petit Robert* 2010-es kiadásában a terminológia a „terminusok vagy olyan szavak és szintagmák szisztematikus tanulmányozása, amelyek célja objektumosztályok és fogalmak leírása (→ lexikográfia); ezen tudományban uralkodó általános elvek"[2]. Ez a leírás már tartalmaz egy taxonomikus és tudományos megközelítést, amely azt hivatott leírni, hogy miként viselkednek a vizsgált nyelvi elemek, azaz a terminusok, amelyek lehetnek egyszerűek és összetettek is. A terminológia tehát egy olyan tudomány, amely a terminusok elméleti hátterével foglalkozik.

A terminológia mint tudományterület Bessé és mtsai (1997) szerint „(1) terminusok, fogalmak és azok viszonyának vizsgálata, (2) terminusok gyűjtésére, leírására, bemutatására alkalmazott eljárások és módszerek összessége, (3) egy meghatározott tárgykör szókincse" (Fóris 2005: 33). A Fóris (2005) által idézett definíció annyiban pontosít a korábbiakhoz képest, hogy a terminológiát mint tudományterületet két különböző részre osztja: az egyik egy szemantikai definíció, amely azt vizsgálja, hogyan

[2] Saját fordítás.

társítható egy fogalomhoz egy terminus, a másik egy „alkalmazott terminológia" meghatározását tartalmazza. Fóris (2005) a korábbi definícióktól eltérően már nem tudományterületet említ, hanem tárgykört. Ez azért lényeges, mert az interdiszciplinaritás elterjedésével már nem köthető a legtöbb terminus egy adott tudományterülethez, hiszen ezeknek egyre gyakrabban van metszetük más szakterületekkel. A Fóris (2005) által említett *magas nyomású folyadék kromatográfia* jól tükrözi ezt a jelenséget: ez a főnévi terminus egyaránt köthető a kémia, az orvostudomány vagy a biológia tudományterületéhez, azonban tárgykör tekintetében csak egyhez tartozik.

Cabré (1999) szerint a terminológia mint tudományág nem újkeletű, a különbség csupán abban van, hogy csak az utóbbi időben vált szisztematikussá, tehát egy kezdetleges állapotból tért át egy teljesen tudományos diszciplínává. Szerinte a terminológia a szakkifejezések vizsgálatával és létrehozásával foglakozó tudományág.

Az, hogy egy fogalom egyaránt jelöli a tudományterületet és az általa vizsgált objektumokat, szintén nem újdonság: más kifejezések esetén is tapasztalhatunk hasonlóakat. A *Magyar nagylexikon* (Vizi 2003, 16. kötet, 796) szerint a szintaxis, azaz a mondattan „a →nyelvtudomány egyik területe, tárgya a →mondat szerkezetének leírása". A meghatározás a mondatot nem szemantikai, hanem szerkezeti, formális egységnek tekinti, azonban ez a lexikon is a fogalom egyik jelentésére fókuszál. A korábban is említett *Petit Robert* (2010) francia egynyelvű szótár alapján a szintaxis a vizsgált elemeket is jelöli, azaz az adott mondat vagy megnyilatkozás „nyelvi egységei közötti viszonyokat"[3].

A fentiek alapján látható, hogy a *terminológia* szó, mind a magyar, mind a külföldi szakirodalomban, többértelművé vált: egyaránt használjuk 'egy adott szakterület szókincse' és a '*terminus technicusok*kal foglalkozó tudományág' jelentésben.

1.2. A terminológia kialakulásának kezdetei

A terminológia kibontakozását a technológia, a szaktudás és a kommunikáció fejlődése eredményezte, így ennek a területnek a születése spontán, mert a szakterületekben folyó terminusegységesítési tevékenység eredménye. E területek gyors fejlődése számos nehézséget okozott, és ez a folyamat már a XVIII. században elkezdődött. A terminológia tudománya így hamar egy nagyon hasznos eszközzé vált, hogy lépést tartsanak ezzel a fejlődéssel (Rey 1995).

[3] Saját fordítás.

A terminológia az idők folyamán számos változáson ment keresztül. A XIX. században a kizárólagos és legfontosabb cél az volt, hogy a lehető legtöbb tudományos területet nemzetközivé tegyenek, valamint az, hogy a terminusok létrehozása és fordítása valamilyen szempontrendszer alapján történjen. Ezidőtájt a legnagyobb hatalommal bíró terminológusok csoportja magukból az adott tudományt művelő tudósokból állt. Ez a XX. században teljesen megváltozott: nyelvészek, mérnökök, különálló terminológusok egyaránt művelik ezt a szakmát (Cabré 1999).

Ezt a tendenciát igazolni látszik az, hogy a modern terminológia atyjának Wüstert, a Bécsi Iskola egyik fő képviselőjét tekintjük, aki végzettsége szerint mérnök volt. A doktori disszertációját is ebben a témában írta 1931-ben (*Internationale Sprachnormung un der Technik, besonders in der Elektrotechnik* 'A szaknyelv, különösen az elektrotechnika nemzetközi szabványosítása'). Ebben amellett érvelt, hogy a terminusokat azonnal standardizálni kell, és meg is tette az első lépéseket ezen cél felé (Picht és Draskau 1985). Wüster számára a szabványosítás volt a legfontosabb cél.

Cabré (2003) is azt állítja, hogy Wüster egész életét a terminológiára áldozta fel. Szerinte Wüster főbb célkitűzései a következők voltak:

1. A szaknyelvekből eltüntetni az összes kétértelműséget, ezzel is elősegítve a hatékony kommunikációt. Számára tulajdonképpen ezt jelentette a terminológia szabványosítása.
2. Meggyőzni a szaknyelvek használóit, hogy miért is előnyös a szabványosított terminológia.
3. Külön tudományágnak nyilváníttatni a terminológiát, és azt tudományos rangra emelni.

Ezeken kívül szerette volna elkülöníteni a terminológiát a nyelvészettől, mert az ő idejében a nyelvészetet a strukturalista megközelítés uralta, ami túlságosan a formális szempontokat tartotta szem előtt, és amelynek a terminológia-szabványosítási folyamathoz nem volt sok köze.

Auger (1988) szerint, amely alapján a terminológia fejlődésének főbb időszakait bemutatjuk majd, a modern terminológia történetében négy nagyobb periódust különül el:

1, a kezdetek (1930–1960)

2, a tudományterület szerveződése (1960–1975)

3, robbanás (1975–1985)

4, terjeszkedés (1985-től)

A terminológia első szakaszát nagymértékben jellemezte a terminusok szisztematikus módszerekkel történő képzése. Wüster munkássága befolyásolta ezt a periódust: ekkor írta a korábban is említett doktori tézisét, amelyben megállapította, hogy a terminológia tudományának célja az, hogy egy olyan hatékony eszköz legyen, amely képes a szakmai kommunikáció során előforduló kétértelműségeket megszüntetni.

A második időszakot azok a felfedezések jellemzik, amelyek az első számítógépek megjelenésével hozhatók összefüggésbe. Ebben az időszakban jelennek meg az első adatbázisok, valamint különböző együttműködési elveket dolgoznak ki a terminológia szabványosítási eljárásaival kapcsolatban.

A harmadik időszakot a terminológiai projektek kitörése jellemzi és a nyelvpolitika fejlődése, ezen belül is különösen a nyelvtervezésé. Ebben az időszakban a terminológia nagy szerepet kapott még a nyelvek modernizációja kapcsán is. Erre és a következő korszakra is igaz, hogy a számítógépek rohamos fejlődése nyomott hagyott ezen a perióduson.

A negyedik szakaszt főleg az informatika jellemzi, ami nagy segítséggel bírt a terminológusok számára is. A számítógépek segítségével a terminológusok (akik nem feltétlenül informatikusok) már olyan programokat használtak, amelyekkel ők is hatékonyan tudtak kezelni hatalmas korpuszokat és adatbázisokat. Számos alkalmazás született, amelyek mindenki számára elérhetővé és kezelhetővé váltak, és hatékonyabbak is lettek, mint az eddigiek. Ráadásul, a természetesnyelv-feldolgozással kapcsolatos különböző területek, például a TE is, ebben a korszakban születtek. A terminológusok már képzettek, mivel a terminológia tudománya már egy különálló és nagyobb érdekérvényesítő hatalommal rendelkezik. Ezen felül, a különböző szakterületek egyre növekvő számú nemzetközi kooperációi és konferenciái egy megfelelőbb, szisztematikusabb és korszerűsített terminológia elterjedését tették lehetővé. Ez az az időszak, amely során a terminológia tudománya felbomlott több alterületre, amelyeket az 1.4. fejezetben ismertetünk.

1.3. A terminológia mint interdiszciplináris terület

Wüster (1981) szerint a terminológia mint diszciplína különböző területek – nyelvészet, logika, ontológia, informatika és egyéb, önálló tudományok – keresztmetszetében áll, mert az előbbi tudomány átvesz fogalmakat az utóbbi tudományokból. Ez azt a következtetést vonná maga után, hogy a több szakterület mezsgyéjén fekvő terminológia nem tekinthető

egy doménnek, mivel az, hogy *interdiszciplináris terület* még nem implikálja azt, hogy ez a diszciplína önmagában is megállja a helyét. Ugyanakkor a terminológia ezen tudományok csak egy részhalmazát használja fel, ezen területek fogalmaiból és elemeiből dolgozza ki a saját koncepcióit, s ezáltal válik külön tudománnyá.

1.3.1. Terminológia és nyelvészet

A terminológiát az alkalmazott nyelvészet egyik kutatási területének tekinthetjük (Cabré 1999). Az alkalmazott nyelvészet célja a nyelvészet felhasználása gyakorlati feladatok megoldására, tehát a nyelvészet elveiből kiindulva próbál azonosítani és megoldani természetes nyelvvel kapcsolatos problémákat. Ebből adódóan az alkalmazott nyelvészet ágai interdiszciplinárisak, ami már az alterületek elnevezéseiből is látszik: első, második és idegen nyelvek tanítása, amely a pedagógia területéhez kapcsolódik, a pszicholingvisztika a pszichológiához, a szociolingvisztika a szociológiához, a nyelvtervezés a politikához.

A terminológia közel áll a nyelvészethez, mert kutatási területe a nyelv, azon belül főleg a szaknyelv lexikai jellemzői. A terminológia tudománya azért kapcsolódik szorosan az alkalmazott nyelvészethez, mert célja a nyelv leírása a szociális funkcióinak tükrében és annak kommunikációs eszközként való kezelése. Egyik feladata egy adott doménhez köthető objektumok, folyamatok megnevezése, amihez a nyelvészetben széles körben ismert eszközöket használ fel, például neologizmusokat (Wüster 1981).

A terminológia a szemantikához is köthető, mert a terminusok célja hasonló, mint a szavaké: kapcsolatot teremteni a nyelv és a való világ között, hiszen a terminusok is a világ elemeit denotálják. A terminológia tudományában gyakran használatos a fogalom (*concept*) elnevezés, amelyek tulajdonképpen kognitív elemek, ebből következően nehezebben tanulmányozhatók. A fogalmak objektumok egy halmazát képviselik, például a *kerékpár* fogalom, amely a „kétkerekű, pedállal rendelkező, lábbal működtethető közlekedési eszköz” képet idézi fel. Ezen objektumosztályok tagjai tehát ugyanazon vagy hasonló, akár tényleges, akár absztrakt tulajdonságokkal rendelkeznek. A konceptumok az azokat jelölő szavaktól vagy terminusoktól függetlenül léteznek, sőt, megjelenésük meg is előzi a nyelvi egységekét, amelyeket önkényesen hoznak létre (Cabré 1999).

1.3.2. Lexikológia és terminológia: két külön terület?

Érdemes megvizsgálni a lexikológia és a lexikográfia, majd a terminológia és az utóbbi kettő közötti különbségeket, mert ezek akár megkérdőjelezhetik a terminológia

létszükségletét. L'Homme (2004) alapján a lexikológia célja a nyelv egy adott területének, lexikájának vizsgálata, valamint egy olyan modell kidolgozása, amely alapján ezen nyelvi komponenst le tudja írni, figyelembe véve a beszélők implicit ismeretét és azon képességét, hogyan hoznak létre új lexikai elemeket az agyban már létező nyelvi struktúrák alapján. A lexikológia célja tehát nem olyan általános elvek kidolgozása, amelyek a szövegegységeket strukturált együttesekbe, például szótárakba szervezik. Ez főleg a lexikográfia célja, amely a lexikológia gyakorlati oldala. Ebből a szempontból felfedezhető párhuzam a lexikológia és a terminológia között: ugyanis mindkettőt elméleti tudományoknak tekintik, amelyeknek a gyakorlati megvalósítással foglalkozó párjai rendre a terminográfia illetve a lexikográfia (Rey 1995). Ezzel kapcsolatban Bergenholtz és Kaufmann (1997) meg is jegyzik, hogy mivel ezek a különbségek nem annyira jelentősek, ezért a szakirodalom a terminográfiát *szakterületi lexikográfiának* vagy *terminológiai lexikográfiának* is nevezi.

Minthogy a terminográfusok a szótárak létrehozására vállalkoznak, ugyanúgy mint a lexikográfusok, felvetődik a kérdés, mi a különbség a terminográfia és a lexikográfia között? Az előbbi a terminusokra koncentrál, míg a lexikográfia sokkal általánosabb elemeket vizsgál, ami elsőre nem tűnik nagy különbségnek, de időnként vannak árnyalatbeli eltérések az erre a célra használt módszerek között (Bergenholtz és Kaufmann 1997).

A lexikológia és a lexikográfia a nyelvi egységeket mindig az előfordulások kontextusának elemzésével végzi, tehát diskurzuselemeknek tekinti őket. Ezzel szemben a terminológia és a terminográfia a terminusok vizsgálatát nem köti annyira a környezethez, mert jelentésüket gyakran ez nem befolyásolja, hiszen, mint ahogy később említjük, a terminusok – elvben – egy adott szakterületen belül egyértelműek. Ezen kívül a terminológiai adatbázisok nem foglalkoznak a bennük szereplő szócikkek szóképzésével és ragozásával (hiszen ez ugyanaz, mint az általános szavak esetében), és nem foglalkoznak a szövegkörnyezettel sem (Cabré 1999).

A lexikológia és a lexikográfia a nyelvnek mind a szinkrón, mind a diakrón aspektusaival foglalkozik, miközben a terminológiát és a terminográfiát csak az előbbi aspektus érdekli, tehát az aktuális használatuk. Terminológiai adatbázisokban ritkán szerepel utalás egy szó létrejöttének elemzésére, sem arra, hogy azt esetleg korábban hogyan használták (Cabré 1999).

A lexikológia – valamint a nyelvészet – nem részesíti előnyben a nyelv fejlődésébe történő emberi, mesterséges beavatkozást, miközben a terminológiának ez a lényege. Miközben arra törekszik, hogy minél inkább szabványosítsa és nemzetközivé tegye a szakterületek terminológiáit, ezzel is azt sugallja, hogy az emberi beavatkozás ezen a területen szükséges. (Bergenholtz és Kaufmann 1997)

L'Homme (2004) szerint az egyik fontos különbség a két tudomány között a fogalomtól a nyelvi jelig tartó út irányában van. Miközben a lexikológia szemasziológiai megközelítéssel él az adatfeldolgozás szempontjából, addig a terminológia onomasziológiai megközelítést alkalmaz. Ez azt jelenti, hogy a lexikológia a szóalakkal kezd, és így jut el a fogalomig, a célja tehát egy adott nyelv szavainak felsorolása, illetve csoportosítása a specifikusságaik alapján, és azok jelentéseinek, esetleges szinonimáinak megadása az adott szövegkörnyezetet figyelembe véve. A terminológia ezzel szemben a fogalomból indul ki, hiszen egyik feladata annak biztosítása, hogy ha egy fogalmat elneveznek, akkor csak annak adjanak nevet, és ne másnak vagy valami hasonlónak. A terminológiai szótárak ezért is adnak nagyon részletes és kimerítő definíciókat minden szakkifejezésre, valamint megadják ezek egymáshoz való viszonyát.

A lexikográfia számára egy anyanyelvi beszélő legtöbbször tökéletesen hiteles, de a terminológiával foglalkozó szakemberek számára csak az adott tudományterület szakértői bizonyulnak megbízható forrásnak.[4]

1.3.3. Terminológia és szabványosítás

Cabré (1999) szerint a terminológia egyik fontos ismertetőjegye az, hogy nagyon fontos szerepet játszik benne a szaknyelvre jellemző terminusok szabványosítása. Ez a sztenderdizálás ugyan a köznyelvben is megvan, azonban a szaknyelvre még inkább jellemző és nélkülözhetetlen, sőt erre kifejezett igény is mutatkozik. Wüster (1931) már szakdolgozatában megmutatta, hogy a terminológia tevékenységének középpontjában a terminusok standardizálásának és normalizálásának kell lennie. A szaknyelvi kommunikációban ezért a nyelvi norma szerepe meghatározóbb, mint a köznyelvben. A terminológiát egy olyan tudományterületnek tekintette, amelynek segítségével a többértelműségek, szinonimák a szaknyelvben kiküszöbölhetők.

A terminológiai szabványosítás folyamatát jól mutatja, hogy a terminológia onomasziológiai megközelítést használ, azaz a fogalomból indul ki, és innen jut el a nyelvi

[4] A lexikográfia és a terminográfia közötti további különbségeket mutatja be Bergenholtz és Kaufmann (1997).

jelig, azaz a terminografikusok célja az, hogy neveket adjanak a már létező fogalmakhoz. Ezért Fóris és mtsai (2009) szerint a terminusok létrehozásakor a legelső lépés a fogalom meghatározása definiálás segítségével. A definíció során meg kell adni azt az osztályt, amelyhez az adott fogalom tartozik, valamint azon egyedi jellemzőket, amelyek azt elkülönítik azon többi elemtől, amelyek ugyanebbe az osztályba tartoznak.

Fóris és mtsai (2009) szerint azt, hogy a terminusok képzése mennyire szabványosított, az is mutatja, hogy már a magyar nyelvújítás folyamán is megfogalmazták azt a három pontot, amely alapján a terminusokat képezni kell: (1) a terminusok létrehozásakor a kiindulópont a jelölt fogalom, (2) egy adott terület terminusainak illeszkednie kell a terminusok rendszerébe, (3) és az újonnan létrehozott terminus nyelvi alakja az adott nyelvtől nem lehet idegen. A (2) pontban említett terminológiai rendszer Fóris (2005) szerint azt jelenti, hogy a terminusok egy adott szakterületen belül egy hálóba rendeződve léteznek, azaz egymásnak alá-, fölé-, mellérendeltjeik. Sőt, ugyanezen tanulmány írja le részletesen, hogyan kell létrehozni új terminusokat, akár más nyelvből történő átvezetés, akár adott nyelvű bevezetés esetében.

Galinski és Nedobity (1988) szerint a szaknyelv feladata tudás, ismeret átadása szakértők között adott és más szakterületen. A szabványosítás a szaknyelvi kommunikáció folyamán elengedhetetlen. Így a kölcsönös megérthetőség szempontjából fontos, hogy a terminológiának preskriptívnek kell lennie, azaz elő kell írnia, hogy adott fogalomra egységesen mely terminust használjuk. Elég csak arra a Fóris és mtsai (2009) által említett példára gondolnunk, hogy a különböző típusú csavarok alkalmazására sokféle termék esetén van szükség, és fontos, hogy az adott elnevezésen mindenki ugyanazt a típust értse. A csavarok fajtáit csoportosíthatjuk a csavarmenet jellemzői segítségével (pl. metrikus vagy *withwort* csavar) vagy a csavarfej alakja alapján (pl. gömbfejű vagy süllyesztett fejű csavar). Ezeket a szakterületek közötti kommunikáció érthetőségére hivatkozva egységesíteni kell.

A terminusok létrehozása azonban nem lehet teljes mértékben preskriptív. Sager (1990) szerint a szabványosítás főbb alapelvei a következők: a normalizálás folyamata (1) egyrészt egyszerűsítés, amely során egy változatot jelölünk ki a többi hátrányára, másrészt (2) csapatmunka, ahol nem a kényszerítés, hanem a megegyezés a lényeges. A sztenderdizálás lépéseit érvekkel kell alátámasztani, hogy azok számára, akiknek használniuk kell, ez mérvadó és követendő legyen.

A köznyelv és a szaknyelv szabványosítása nem teljesen azonos. Cabré (1999) szerint a terminológiai sztenderdizálás többrétű: nemzetközi, nemzeti és regionális szinten is történhet, amelyek közül a legelsőt emeli ki részletesebben. Ezzel szemben a köznyelvben nemzetközi szinten ritkán történik szabványosítás. A nemzetközi szabványosításnak két fő szervezete van: IEC (International Electrotechnical Commission – Nemzetközi elektrotechnikai bizottság), amelyen belül 1910 óta létezik egy terminológiai bizottság. A másik ilyen az ISO (International Organization for Standardisation – Nemzetközi szabványosítási szervezet), amely az elektrotechnikai és elektronikai terület által le nem fedett terminusokat szabványosítja (Cabré 1999).

A terminusok szabványosítása sokkal több műveletet von maga után, mint a köznyelv sztenderdizálása. Az előbbi kategóriába tartozik a fogalmak egyesítése, új terminusok alkotása, rövidítések rögzítése, terminusok definícióinak meghatározása, pontosítása, a szinonimák számának csökkentése a homonimák eliminálásával egy időben (Cabré 1999).

Sager (1990) szerint mind a köznyelv, mind a szaknyelv esetében neologizmusok segítségével képezhetünk új szóalakokat a nyelvi szabványosítás folyamatában. A kettő között azonban vannak eltérések: (1) a köznyelvi neologizmusok spontánabbak, és általában rövidebb élettartammal rendelkeznek, azaz a szaknyelvi neologizmusok a megnevezés kényszere miatt jönnek létre, és tovább maradnak is fenn. (2) A köznyelvi neologizmusokat nem érinti a szinonímia jelensége, mert attól függetlenül léteznek, hogy arra a fogalomra létezik-e már szó, esetleg stílusbeli különbség lehet a kifejezések között. Ezzel szemben a szaknyelvi terminusok elutasítják szinonimákat, mert azok jelenléte nem hatékony. (3) A köznyelvi neologizmusok forrása lehet például regionalizmus, szociolektus vagy jövevényszó, nem jellemző rájuk a szaknyelvben használatos latin vagy görög előtag. (4) A köznyelvi neologizmusok egy adott nyelven belül jönnek létre, a szaknyelvieket pedig nemzetközi használatra tervezik.

Így azt állapíthatjuk meg, hogy a terminológia esetében, a lexikológiával ellentétben, a terminusok szabványosítása különösen fontos, és erre igény is van. A köznyelvi egységek esetében is van természetesen szabványosítás, de az ennél kisebb mértékű. A terminológia ezért preskriptív, de akkor hatékony, ha a szakértők véleményét is figyelembe veszi. Ha ezt nem teszi meg, sok esetben a szaknyelv művelői más alakot fogadnak el és kezdenek használni. Zimányi (2006) szerint például az idegen terminusok magyarosítása sem mindig marad fenn, ilyen például a *file*, amelyet *állomány*ként

próbáltak magyarosítani, végül pedig maradt az idegen szó, csak *fájl* alakban. Ezzel ellentétben sok esetben, például a labdarúgás terminológiájában, a legtöbb kifejezést sikerült magyarrá formálni, például az *offside* helyett ma már *les*t használunk.

1.3.4. A terminológia kapcsolata más tudományterületekkel

A terminológia egyaránt kapcsolatban áll az ontológia tudományával, annak főként számítógépes nyelvészeti vonatkozásában (Jacquemin és Bourrigault 2003). A klasszikus filozófiai értelemben az ontológia célja egyedek csoportosítása különböző nem- és fajkategóriák alapján. Valójában ennek számítógépes nyelvészeti megközelítése sem különbözik ettől: az ontológia végtermékei lehetnek fogalmi hálók, lexikonok, szótárak, valamint *thesaurus*ok (számítógépes alkalmazásokhoz készülő szinonimaszótárak). Az ontológia célja tehát egy adott fogalmi terület feltérképezése: a végtermék tartalmazza az adott terület terminusait, az azok közötti viszonyokat, vagyis az adott domén fogalmi rendszerét, amelyekben minden hierarchiába rendezett (Vossen 2003).

A terminusok között az egyik ilyen viszony az *IS-A* (’az egy’) kapcsolat, amely egy adott egyed hiperonimáját vagy hiponimáját adja meg. Minden nyelvben észrevehető, hogy az ilyen típusú szerkezetekben az *IS-A* bal oldalán található a tőle jobbra álló entitás hiponimája. Erre példa az a mondat, hogy *a penguin* ***is a*** *bird*, azaz *a pingvin* ***az egy*** *madár*, amely fordítva már nem lenne helyes, tehát *a bird is a penguin* vagy *a madár az egy pingvin* jelentésüket tekintve már nem elfogadható mondatok. A másik gyakran vizsgált formula a *HAS-A*, amelyben a kifejezéstől balra álló entitást tartalmazza a kifejezés jobb oldala. Erre példa az a szerkezet, hogy *a penguin has a beak*, azaz *a pingvinnek van csőre*. Természetesen egy ontológiában másfajta viszonyokat is feltérképezhetünk, például hogy mely elemnek mik a szinonimái, meronimái stb. Mindezt általában egy irányított gráf formájában ábrázolják, ahol a kapcsolatban álló egyedek és attribútumok között irányított él van, és az él címkéje a viszonyt tartalmazza (Vossen 2003).

Az ontológiákat lehet a terminológia szolgálatába állítani, ekkor a gráfban az egyedek terminusok (vagy az azok által denotált konceptumok), és a köztük lévő élek a köztük fennálló viszonyt írják le. Például az informatika területén használt *háttértároló* hiponimája a *merevlemez* vagy a *CD-ROM* amelyek közül az első *HAS-A* viszonyban van az *olvasófej* terminussal, és attribútuma az *olvasható* tulajdonság, míg a másodiknál ez a kettő nincs meg. A 5. fejezetben is említeni fogjuk, hogy a terminológiával kapcsolatos

tudományterületek nem állnak mindig kapcsolatban az ontológiával: sok terminológiakivonatoló nem foglalkozik ontológiai háló létrehozásával, csak a terminusok kinyerésével.

Másik kapcsolódó terület az informatika, amellyel a terminológia kapcsolata bilaterális: a terminológia feladata az informatika tudományát ellátni a megfelelő terminusokkal és fogalmi hálókkal, ugyanakkor az előbbi elég erősen támaszkodik is rá, mert a terminológia számára lehetővé teszi a terminusok automatikus kezelését. Ez a folyamat egyre kevesebb emberi beavatkozást igényel, ami a munkát felgyorsító nagyobb teljesítményű hardvereknek és a mesterséges intelligencia fejlődésének köszönhetően egyre jobban és hatékonyabban felváltja a terminusokkal kapcsolatos monoton tevékenységeket, például a terminológiakivonatolást (Cabré 1999).

1.4. A terminológia különböző területekre történő tagozódása

Ahogy a terminusok kezeléséhez egyre több feladat társult (pl. sztenderdizálás, terminológiai szótárak létrehozása, automatikus terminológiakivonatolás), a terminológia kutatási területe is egyre nagyobbá vált, így szükséges volt az addig a terminológia tudományterületéhez tartozó tevékenységek egy részét más tudományterületbe helyezni. Az első nagy lépés azzal kezdődött, amikor a terminológiából létrehozták a terminográfiát (Rey 1979). A terminográfia célja a terminusok kikutatása, kezelése, létrehozása, míg a terminológia feladata a terminusok használata által felvetett kérdések megválaszolása, tehát egy olyan konceptuális keretet ad, amelynek segítségével a terminusok a hozzájuk tartozó fogalmakkal összekapcsolhatók. Természetesen a két terület nem választható szét ilyen egyszerűen egymástól, mert a fejlődésük végett mindkettőnek szüksége van a másikra.

Fóris (2007) szerint a terminológiai paradigmaváltás az 1990-es években következett be, aminek oka elsősorban az, hogy a technika, a gazdaság egyre gyorsabb ütemben kezdett el fejlődni. Ezáltal (1) egyre nagyobb mennyiségű új fogalom keletkezik, amelyeknek természetesen nevet kell adni, és ha lehet, ezt minél korábban, mivel ezek a nyelvbe is gyorsan beépülnek. Ez azt vonja maga után, hogy (2) mivel a terminusok, új szavak gyorsan elterjednek a nyelvben, ezért a későbbiekben már nehezebb ezeket korrigálni. Ezért az újonnan megjelenő fogalmakra minél előbb meg kell találni a megfelelő terminust. (3) A gyorsan vagy nem megfontolt módon kialakított terminusalkotás során a megválasztott új szó sokszor helytelen, amely esetleg zavart okoz

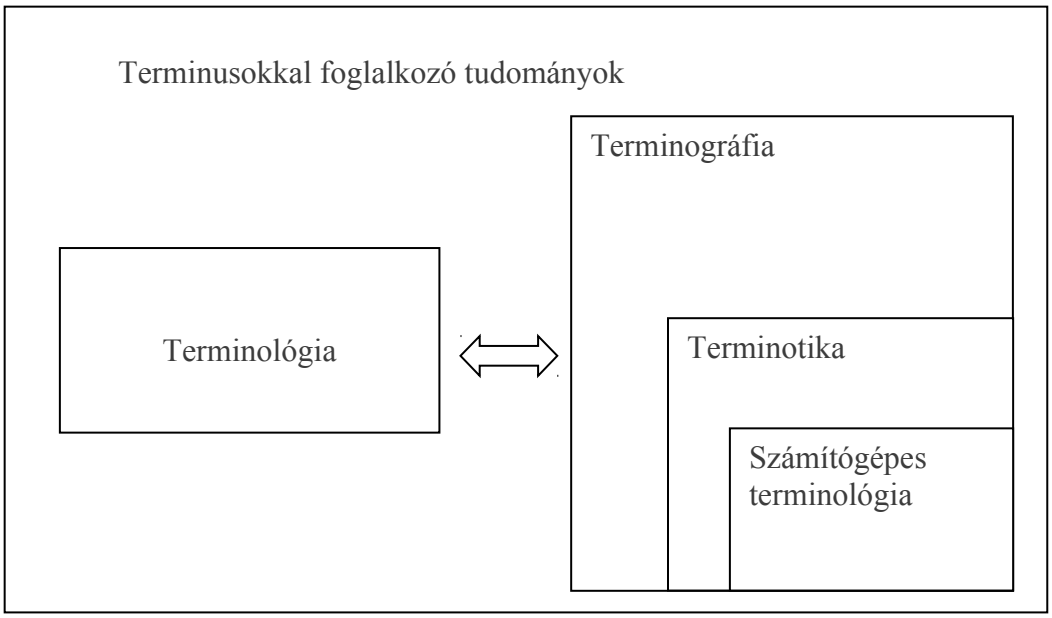
a nyelvhasználatban és egy adott terület fogalomrendszerében. Bizonyított, hogy ha a véletlenszerű megnevezések, jobb híján, teljesen elterjednek a nyelvben, akkor ezt már nagyon nehéz korrigálni. Gyakran előfordul az is, hogy a későbbiekben mindkét variáns megmarad: például az informatika területén gyakran mind a magyarított, mind az eredeti alak is használatos. Erre példánk lehet a *merevlemez* fogalom, amely az informatika hőskorában ebben a formában még nem létezett, csak az angol változata, a *HDD* vagy a *winchester*, amit ma már *vincseszter* alakban kell írnunk (Laczkó és Martonfi 2004: 1467). A *router* szóra létező *útválasztó* magyarosított, és így szabványszerű kifejezést valószínűleg kevesen ismerik, mert Várlaki (2005) szerint a *router* magyarul is 'router', azonban Iványi (2006) szerint csak a magyarosított változat a helyes. (4) Mivel a tudományos megközelítés ma már mindenhol fontos, így ez kihat a terminológia fejlesztésére is, ezért a terminológiarendszert is inkább csak szakterületi ismeretekkel és nyelvészeti tudással lehet jól kialakítani. (5) Napjainkban szintén elengedhetetlen, hogy a terminusok megfelelően legyenek definiálva, mert ez biztosítja, hogy ezeket ellentmondások nélkül lehet használni. Hiába a megfelelően létrehozott terminus, ha az nem feleltethető meg egyértelműen egy fogalomnak, ezért (6) nagyon fontos, hogy a szakterületek képviselői részt vegyenek a terminusalkotási folyamatban.

Az is igazolja azt, hogy a terminológia különálló diszciplína, hogy tárgyai elsősorban nem maguk a terminológiai egységek, hanem ezek kapcsolata az általuk denotált fogalmakkal. Mint korábban is említettük, a terminológia főleg onomasziológiai megközelítést használ, tehát a vizsgálatok kiindulópontja mindig a fogalom, és abból vezetik le a jelet. Észrevehető még az is, hogy a terminusok nagyobbrészt egymásnak megfeleltethetőek a különböző nyelvekben, tehát a főbb vizsgálódási szempontnak nem a jelölőknek kell lenniük, hanem a jelölteknek. Mindez azt mutatja, hogy a terminológia tudománya egy teljesen más és határozott irányba indult, főként a terminológia szabványosítása felé (Cabré 2003).

Minthogy a jelen disszertáció a terminológiának főleg az informatikai aspektusaival foglalkozik, ezért azon tudományágakat is célszerű megvizsgálni, amelyek az informatika és a terminológia viszonyából alakulnak ki. Ilyen tudományterület a terminotika és a számítógépes terminológia. A terminotika a terminusok kezelése, gyűjtése során az összes olyan tevékenységet takarja, amely során valamilyen számítógépes alkalmazást használunk. Mivel manapság mind a terminotika, mind a terminográfia használ számítógépet valamely folyamatában, ezért első látásra nem lenne érdemes őket

szétválasztani. A különbség csupán annyi, hogy a terminotika kifejezés azt takarja, hogy ez utóbbi nagyobb fontosságot rendel a számítógépes alkalmazásokhoz a terminológiakezelés során. Ennél fontosabb különbség a terminográfia és a számítógépes terminológia között van. Az utóbbi megjelenése egyértelműen a természetesnyelv-feldolgozást megvalósító új tudományterületek (pl. információkinyerés, jelentés-egyértelműsítés) megjelenésével indokolható. Ez magyarázza magának a számítógépes nyelvészet tudományának a létrejöttét, majd azon belül a számítógépes terminológiáét. A számítógépes terminológia a természetesnyelv-feldolgozás egyik alterülete, amely matematikai, mesterséges intelligenciához kapcsolódó algoritmusokat használ fel, hogy azokat szakszövegeken alkalmazza, így foglalkozik a terminusok automatikus kinyerésével, azok rendszerezésével stb. (L'Homme 2004).

Az 1.1. ábra szemlélteti a terminusokkal foglalkozó különböző tudományterületek kapcsolatát. A terminusokkal foglalkozó tudományterületek két fő ágra tagozódnak: az egyik a terminológia, a másik a terminográfia. Mint ahogy az ábrán is látszik, a kettő között ugyan nincs átfedés, de a továbbfejlődéshez mindkettőnek szüksége van a másik eredményeire (Rey 1979). A terminográfián belül a terminotika, majd azon belül a számítógépes terminológia áll. Ez utóbbi sokat merít a mesterséges intelligencia területéről, de ez az áramlás csak egyirányú, mert a mesterséges intelligencia nem tartozik bele a klasszikus értelemben vett terminológia területébe.



1.1. ábra: A terminusokkal foglalkozó tudományterületek kapcsolata

1.5. A disszertáció területe

A TE a terminológia mint tudomány fogalmába nem illeszkedik bele: nem foglalkozik a terminusok szemantikai aspektusaival, tehát a terminusok és a fogalmak (*concepts*) viszonyával, a terminusok létrehozásának elméleti kérdéseivel, valamint azok szabványosításával. A TE-t így az elméleti jellegű terminológia nem fedi le, mert gyakorlati feladatot old meg, ezért ebből a szempontból a terminográfia területéhez köthető. Mivel a terminológiakivonatoló alkalmazások kimenetét gyakran használják fel terminológiai adatbázisok létrehozására, ezért a terminográfia területére is tehető, ugyanakkor a TE a terminográfiai munkának csak a terminusok kigyűjtésével foglalkozó szakaszát valósítja meg automatikusan.

A terminotika tudományterületének definíciója (számítógép használatát igénylő terminográfiai munka) egyértelműen vonatkozik az automatikus terminológiakivonatolásra is. A terminusok automatikus kinyerése csakis számítógép segítségével történhet, ami egyértelműen igazolja a disszertáció területének a terminotikához való tartozását.

Az automatikus terminológiakivonatolás elsősorban a természetesnyelv-feldolgozás témaköréhez köthető, amely a számítógépes nyelvészet egyik alterülete (Jacquemin és Bourrigault 2003). A vizsgálatunkban gyakran alkalmazunk matematikai vagy mesterséges intelligenciából ismert algoritmusokat terminusok kinyerésére illetve szűrésére, és az alkalmazásunk szempontjából fontosabb szerepet töltenek be az algoritmusok, mint a terminológiai adatbázisok létrehozása, így a doktori disszertáció területe a számítógépes terminológia is.

2. A szaknyelv jellemzői

E fejezet célja a szaknyelv bemutatása, valamint a szaknyelv és a köznyelv viszonyának vizsgálata. Erre egyrészről azért van szükség, mert a terminusok definíciójához, amelyet a 3. fejezetben írunk le bővebben, elengedhetetlen a szaknyelvek definíciója, hiszen a terminusokat egyrészről az határozza meg, hogy szaknyelvi szövegekben szerepelnek. Másrészről a szakszöveg fogalmának ismeretére a vizsgálatunk korpuszának meghatározásához van szükség: mivel a TE-hez terminusok kellenek, valamint a terminusok szakszövegekben találhatók, ezért a korpusznak valamilyen szakterülethez kötődő szakszövegnek kell lennie.

A szaknyelvet legtöbb esetben a köznyelvhez való viszonya és a kettő közötti különbségek alapján definiálják (pl. Gotti 2003), azaz nincs olyan definíció, amelyben a köznyelv fogalma ne szerepelne. A jelen fejezetben ezeket a különbségeket vesszük górcső alá (2.1. fejezet), majd a szaknyelv konkrét jellemzőit, amelyek lehetnek szintaktikai vagy lexikális szintűek, mindezt egy példán alátámasztva (2.2. fejezet).

2.1. A szaknyelv viszonya más nyelvi rétegekkel

Beaugrande (1987) összefoglalja, milyen megközelítéssel lehet a szaknyelvet a köznyelvvel összehasonlítani. Két fő megközelítés létezik: az egyik szerint a szaknyelv egy teljesen más kód, mint a köznyelv (pl. Kiss 1995), a másik szerint a szaknyelv a köznyelv egy változata (pl. Sager és mtsai 1980, Gotti 2003). Ez utóbbi az elfogadott nézet, mert a szaknyelvet egyben tekinthetjük olyan változatnak is, mint a szociolektusokat vagy a dialektusokat (Cabré 1999).

Cabré (1999) szerint a szaknyelv a köznyelvnek egy alfaja, alkódja. A köznyelv rendelkezik egy szabálykészlettel a benne szereplő elemekre vonatkozóan, ezen szabályokat jelöletlennek nevezzük. Ezzel szemben a szaknyelv olyan alkódokkal rendelkezik, amelyek többé-kevésbé átfedést mutatnak a köznyelvvel, de azokat esetlegesen kibővítik, vagy választás esetén vannak olyan lehetőségek, amelyeket a szaknyelv előnyben részesít, például a személytelen szerkezeteket.

Sager és mtsai (1980) nem választja szét a kettőt, mert úgy véli, hogy a határ nem egyértelmű, tehát a kettő közötti különbség inkább csak fokokban mérhető: az, hogy a szaknyelvben a köznyelvi jellemzők mennyire minimalizáltak. A köznyelv használata

kevésbé tudatos, azonban a szaknyelv használatában a köznyelvben elismert spontaneitás kisebb mértékben van jelen.

Gotti (2003) ugyancsak úgy vélekedik, hogy a szaknyelv nem egy szociolingvisztikai változata a köznyelvnek, hanem egy olyan nyelv, amely specifikus jegyek összességével rendelkezik. Ezek számottevőleg többet fordulnak elő együtt a szaknyelvben, mint a köznyelvben. Legjelentősebb és legpertinensebb különbség a lexikában van, ugyanis a szaknyelv – jellegéből adódóan – több szakkifejezést használ, mint a köznyelv.

A matematikából ismert halmazok fogalmával is szemléltethető a két nyelvréteg közötti különbség. Kocourek (1982) szerint a nyelvi változatokat tekinthetjük olyan halmazoknak, amelyek egymással kapcsolatban állnak és átfedik egymást. Az összes halmaz metszetében található a köznyelv, hiszen minden egyéb nyelvi réteg (pl. argó, szaknyelv) erre építkezik. A halmazok egy része felfogható a különböző szaknyelvek egy-egy reprezentációjaként.

Kiss (1995) a magyar nyelvre vonatkozóan azt állítja, hogy az három főbb szövegtípusban jelenhet meg: vannak társadalmi, köznyelvi és területi változatok. Ezen kategorizáláson belül a szaknyelvet a társadalmi nyelvhasználat kategóriájába sorolja a csoportnyelvekkel együtt, azaz egy teljesen külön kategóriába a köznyelven kívül.

A szaknyelv első főbb ismertetőjegye az, hogy elsősorban bizonyos szövegtípusokat alkalmaz. Rendelkezik bizonyos nyelvészeti jellemzőkkel: (1) a szókészlet, azaz olyan nyelvi egységeket is tartalmaz, amelyek csakis azon a területen használatosak, vagy ha más területen is használják azokat, akkor ott eltérő a jelentésük. A szakszövegek ezen kívül (2) szabatos megfogalmazásúak, tömörebbek és rendszerezettebbek, és (3) általában speciális ismereteket továbbítanak. Ezenkívül, a szaknyelvben az információ közlése rendszerezett, vagyis a szöveg struktúráját olyan gráfként kezeljük, amelyben a csomópontokban a kifejezések találhatók, és az azok közötti kapcsolatok adják az ismeretanyagot (Cabré 2003).

Kocourek (1982) szerint a szaknyelv mint kommunikációs eszköz definiálása nemcsak a természetes nyelvhez képest történhet, hanem a szemiotikán keresztül is. Ez alapján a szaknyelv olyan információtovábbító és -cserélő rendszer, amely egyszerre több nyelvi alkódot használ, és ebben legnagyobb részben a természetes nyelv található. Ám ennél sokkal több kódot is alkalmaz, hiszen a szaknyelvek esetében nemcsak a nyelvet

használjuk, hanem különböző háromdimenziós (pl. modellek), kétdimenziós (pl. térkép, ábra) vagy szimbolikus elemeket (pl. szimbólumtáblázat) is.

A klasszikus Jakobson-elmélet (Jakobson 1963) szerint a nyelvnek hat funkciója van: referenciális, érzelmi, kapcsolattartó, kapcsolatteremtő, játékos és metanyelvi. Ezek az általános nyelvhasználatban mind előfordulnak, bár nem azonos súllyal. A szaknyelvet viszont csak a referenciális funkció jellemzi, mert fő célja ismeretek közvetítése egy adott területen belül. Kocourek (1982) szerint fontos azt is hangsúlyozni, hogy a szaknyelv főleg kijelentő mondatokkal operál, ami abból adódik, hogy a szaknyelvben az érzelmi funkció nem domináns.

Kocourek (1982) azt a lehetőséget is felveti, hogy a szaknyelv a mesterséges nyelvek egy változata, hiszen a kettő számos tulajdonsága megegyezik. A mesterséges nyelvekhez hasonlóan a szaknyelv sem enged meg – elvileg – poliszémiát, és olyan új szó létrehozását, amelyet még nem definiáltak pontosan – igaz, az utóbbira ezek csak elméletben vonatkoznak. A szaknyelvben előforduló terminusok nemzetközi szinten érvényesek. A szaknyelv is csak korlátozott funkciókra és környezetben használható. Fontos hasonlóság még, hogy mindkettő megenged emberi beavatkozást, de ez a megengedés a mesterséges nyelv esetében minimális. Mindez azt mutatja, hogy a szaknyelvek rendelkeznek a mesterséges nyelvek egyes tulajdonságaival, azonban a különbségek száma és jelentősége sokkal nagyobb, így nem célszerű sem a mesterséges nyelvet, sem a szaknyelvet a másik egy alfajának tekinteni.

2.2. A szaknyelv lexikai és szintaktikai jellemzői

A nyelvi jellemzők bemutatásakor nem térünk ki a fonológiai és morfológiai jellemzőkre, mert a szaknyelvben ezek a jegyek nem pertinensek. A szaknyelvek lexikai jellemzői a terminusok, ezekre a 3. fejezetben térünk ki bővebben. A szintaktikai jegyek bemutatásához az alábbi szabadalomrészlet adhat példát, amely annak angol és francia nyelvű összefoglalóját tartalmazza, valamint egy szemléltető ábrát is, amely a szabadalom absztraktjának melléklete.

(EN) The invention relates to a technique for connecting a node (MN) to a subnetwork (SN1, SN2) by way of an access point (PA1-PA3) in a communication network (RC), implemented by the node. A method of connection comprises: - a step of selecting a new access point from among a plurality of detected access points, - a step of detecting a subnetwork modification, characterized in that it furthermore comprises a step of receiving by the node an item of information broadcast by the new access point, said item of information being representative of a subnetwork to which the new access point is connected and in which the step of detecting a subnetwork modification is performed by means of the item of information received.

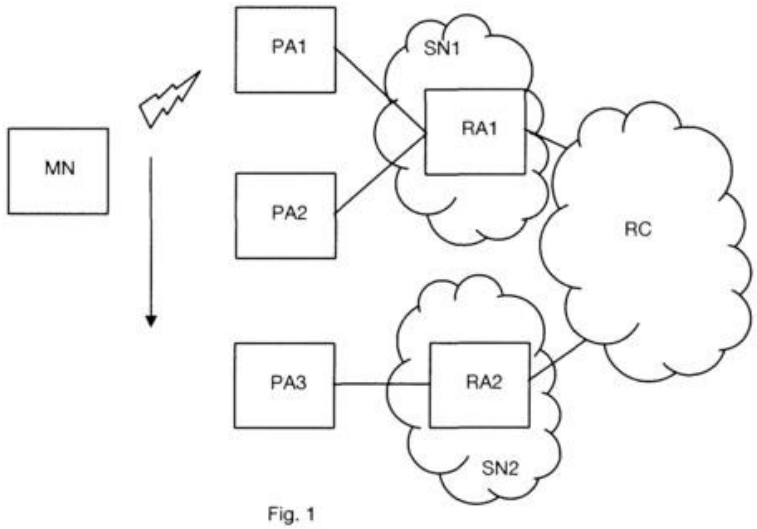


Fig. 1

(FR) L'invention concerne une technique de connexion d'un nœud (MN) à un sous-réseau (SN1, SN2) par l'intermédiaire d'un point d'accès (PA1-PA3) dans un réseau de communication (RC), mis en œuvre par le nœud. Un procédé de connexion comprend : une étape de sélection d'un nouveau point d'accès parmi une pluralité de points d'accès détectés ; une étape de détection d'une modification de sous-réseau ; caractérisé en ce qu'il comprend en outre une étape de réception par le nœud d'une information diffusée par le nouveau point d'accès, ladite information étant représentative d'un sous-réseau auquel le nouveau point d'accès est connecté et dans lequel, l'étape de détection d'une modification de sous-réseau est effectuée au moyen de l'information reçue. [5]

Cabré (1999) szerint a szaknyelv fő jellemzője a lexikája, ugyanis ez az a terület, amelyben a legjobban eltér az általános nyelvhasználattól. Egy adott szaknyelvi szöveg abban

[5] PCT/FR2009/050505 számú szabadalom leírása, amelynek címe: (EN) TECHNIQUE FOR CONNECTING A NODE TO A SUBNETWORK BY WAY OF AN ACCESS POINT IN A COMMUNICATION NETWORK (FR) TECHNIQUE DE CONNEXION D'UN NOEUD À UN SOUS-RÉSEAU PAR L'INTERMÉDIAIRE D'UN POINT D'ACCÈS DANS UN RÉSEAU DE COMMUNICATION http://www.wipo.int/pctdb/en/wo.jsp?WO=2009125125
Saját fordítása:
A találmány egy olyan technikához kapcsolódik, amelynek során egy csomópontot (MN) összekapcsolunk egy alhálózattal (SN1, SN2) egy elérési pont (PA1-PA3) segítségével egy a csomópont által megvalósított kommunikációs hálózatban (RC). A csatlakozás módszere az alábbi lépéseket tartalmazza: - egy új elérési pont kiválasztása az észlelt elérési pontok összessége közül, - alhálózat változásának észlelése, amelynek jellemzője, hogy még ezen kívül tartalmazza az új elérési pont által szórt adatelem csomópont általi kézhezvételének folyamatát, amely adatelem egy olyan alhálózatot jellemez, amelyhez az új elérési pont csatlakoztatva van, és amelyben az alhálózat változásának észlelése a kapott adatelem segítségével történik.

különbözik egy általános szövegtől, hogy sok olyan szót vagy többszavas egységet használ, amely az adott doménra vonatkozik. A fenti példában ezek az elemek a *subnetwork* 'alhálózat'*, item of information* 'adatelem'.

A szaknyelv egyik legszembetűnőbb szintaktikai jegye a személytelen szerkezetek használata. Ez számtalan formában valósulhat meg: szenvedő szerkezetek, főnévi szócsoportok gyakori használata.

A szenvedő szerkezeteket a személytelenség kifejezésére használja a szaknyelv, amelynek a személytelenségen kívül más funkciója is van: a szignifikánsabb jelentést tartalmazó elemeket a mondat elejére tenni (Strevens 1977, Gotti 2003). Erre példa a szövegben az utolsó mondat végén található, ahol a szenvedő szerkezet nemcsak a személytelenségre utal, hanem a mondat szerkezetét is könnyebben átláthatóvá teszi:

> […]a subnetwork to which the new access point is connected and in which the step of detecting a subnetwork modification is performed by means of the item of information received

A mondat hosszúságát tekintve többféle nézet is uralkodik a szakszövegekkel foglalkozó szakirodalomban. Gotti (2003) és Strevens (1977) szerint a szaknyelvre a köznyelvben előforduló mondatokhoz képest hosszabb mondatok jellemzőek többszörös beágyazással és alá-, illetve mellérendeléssel. Kocourek (1982) ezzel szemben azt állítja, hogy a szaknyelvekben a mondatok rövidebbek. Ezért azt a megkötést kell tennünk, hogy a mondatok hossza nagyban függ attól, milyen szakterületről és milyen típusú, célú szövegről van szó. Gotti (2003) szerint a jogi szaknyelvben a többszörösen összetett mondatok száma jelentősebb az általános nyelvhasználathoz képest. Ez a tendencia ezen a szakterületen gyakran annyira erős, hogy a pontos megfogalmazás kritériuma teljesen elnyomja az átláthatóság kritériumát. A szabadalmak is hasonlóak ebből a szempontból, hiszen a kötelezően egy- vagy kétmondatos főigénypontokban vagy összefoglalókban az adott felfedezés részleteit ebbe, illetve ezekbe a korlátozott szintaktikai egységbe kell foglalni. Erre jó példa a fenti szabadalmi összefoglaló, amelyben a pontos megfogalmazás többszörösen összetett mondatokat és ezáltal kevésbé átláthatóbb szerkezetet von maga után.

A szaknyelv egy másik jellemzője a főnévi csoportok gyakori használata, aminek célja a személytelenítés és az objektivitás kifejezése. A szaknyelvben előforduló főnévi csoportok általában hosszabbak is, mint a köznyelvben, ami annak köszönhető, hogy az összetett elő- és utómódosítói szerkezetek segítenek az objektivitás és a pontosság előmozdításában. A vonatkozói és egyéb alárendelt mondatokat pedig igyekeznek tömör,

hosszú főnévi csoportokkal felváltani (Kocourek 1982, Gotti 2003). A fenti példában az *an item of information broadcast by the new access point* lehet kiindulási pont, ahol a főnévi fej után álló harmadik alakú ige a *that is broadcast* vonatkozói mellékmondat alak helyett áll.

A szaknyelvekben az igeidők és az igemódok használata is korlátozódhat (Cabré 1999). A szaknyelv pragmatikai céljainak legjobban a kijelentő mód jelen idő felel meg, ezért ez is terjedt el a legjobban (a fenti példában is csak ez az igeidő szerepel). Ezenfelül más igeidők és -módok is előfordulhatnak, de az már nagyban függ az adott szakszöveg céljától, például egy használati útmutató leírásában a felszólító mód is előfordulhat. Az angolban a technikai leírásokban előfordulhat felszólító mód, de a francia nyelvre ez nem jellemző, és ehelyett a mód helyett, a főnévi igenevet mint ragozatlan alakot használják, ezzel is kerülve a direkt felszólítást.

A szaknyelvekben észrevehető egy tendencia az egyes számú személyes névmások használatában (ha a szenvedő szerkezet nem használható): a szakszövegekben feltűnően kerülik az egyes szám első személyű névmást a többes szám első személyű személyes névmás előnyére. Ezt a folyamatot nevezik személycserének, amelynek a szóban forgó *én ~ mi* alternáció a leggyakoribb formája. Jobst (2007) szerint itt nem a hatalom kifejezésére szolgáló királyi többesről van szó, hanem a szerénység kifejezésére szolgáló *mi* alakról. Ez a tendencia a magyar, az angol és a francia nyelvre is igaz, az utóbbinál az elnevezése *nous de modestie* 'szerénységi *mi*'(Kelemen 2001).

3. A terminusok jellemzői

A 3. fejezet célja a terminusok meghatározására, valamint jellemzőinek kimerítő felsorolására tett kísérletek bemutatása. A *kísérlet* szó használata jelen esetben azzal indokolható, hogy a terminológiai szakirodalomban szereplő látszólag egyszerű definíciók mögött megannyi buktató rejtőzik, és egyik sem tartalmaz elegendő információt a terminusok egyértelmű azonosítására. Mindamellett szükség van a terminus minél pontosabb definiálására, mert a terminuskinyerési folyamat során szükséges a kivonatolt terminusok kézi ellenőrzése.

A 3.1. fejezetben ismertetjük a terminus klasszikus definícióját, majd azt fejtjük ki bővebben a benne szereplő fogalmak pontosításával. A 3.2. fejezetben további definíciókról esik szó, ami már mutatja, hogy a klasszikus definíció sok tekintetben pontosításra szorul. A 3.3. fejezet célja a terminusok köznyelvi egységekkel történő összehasonlítása, amely szintén elősegíti a terminuslista szűrését a terminuskinyerési folyamatban. A 3.4. fejezetben bemutatjuk, hogyan definiálják a TE-vel kapcsolatos cikkek a terminusokat. A 3.5. fejezetben azt írjuk le, hogy a terminusoknak milyen morfoszintaktikai jellemzőik vannak, mert ezek elősegíthetik a dolgozat gyakorlati részének megvalósítását. A végső részben (3.6.) leírjuk, hogy az addigiak alapján a disszertációban hogyan definiáljuk a terminust.

A terminusok szintaktikai kategóriájuk szerint több csoportba sorolhatók: léteznek igei, melléknévi, főnévi és határozói terminusok. Mivel a saját (és a legtöbb) terminológiakivonatoló kizárólag főnévi terminusokat nyer ki, így a jelen fejezetben a példák nagy része főnévi terminus, de az ebben a részben szereplő legtöbb definíció bármilyen típusú szakkifejezésre alkalmazható. Egyetlen kivétel a terminusok összetettségét taglaló 3.5. fejezet, amely csak a főnévi terminusokra érvényes megállapításokat tartalmaz.

3.1. A terminus klasszikus definíciója

A klasszikus terminológiai definíció megalkotása Wüster (1976, 1981) művéhez köthető. Ez az a nézet, amelyet a terminológiai közösség elismer és elfogad, amikor a terminus meghatározására kerül sor (Petit 2001). E szerint egy lexikai egység terminusi mivolta a következő három feltételhez köthető: (1) a terminus kapcsolódik egy (és csakis egy) fogalomhoz, (2) megnevezi ezt a fogalmat és (3) valamilyen szakterülethez köthető.

A feltételrendszer alapján – elvben – a terminusok egy és csakis egy fogalomhoz tartoznak, de ugyanez a másik irányban is igaz, azaz egy fogalmat csak egy terminus nevezhet meg. Mint ahogy az 1.3.2. fejezetben leírtuk, a terminológia tudománya onomasziológiai megközelítést alkalmaz, azaz a fogalomból indul ki, hiszen egyik feladata annak biztosítása, hogy ha egy fogalmat elneveznek, akkor azt elnevezzék, és csak annak a fogalomnak adjanak nevet, ne másnak vagy valami hasonlónak, így elvben minden fogalomnak rendelkeznie kell névvel. Ezen tények alapján a fogalmak és a terminusok halmaza közötti viszony bijektív leképezés. A terminus jellegéből adódóan csak akkor terminus, ha az egy szaknyelvi szövegben szerepel.

A fenti definíció azonban több kérdést is felvet: hogyan határozható meg a fogalom és hogyan a szakterület fogalma? A következő alfejezetben ezen kérdésekre adunk választ.

3.1.1. A fogalom definíciója

Wüster (1976, 1981) szerint a fogalom (*concept*) egy mentális szerkezet, amely egy többé-kevésbé önkényes absztrakció segítségével jön létre, és amelynek célja, hogy a belső vagy külső világ konkrét objektumait osztályozza. Ez azt jelenti, hogy a fogalmak nem feltétlenül feleltethetők meg egy-egy valós objektumnak, csak azok egy csoportjának. Erre példa az első fejezetben már említett *kerékpár* szóval jelölt fogalom, amely legtöbbször ’kétkerekű, pedállal rendelkező, lábbal működtethető közlekedési eszköz’ mentális konstrukciót idézi fel. Ebből az derül ki, hogy a fogalmakat tulajdonságok összességei alkotják.

Cabré (1999) szerint ezek a tulajdonságok két csoportra oszlanak: az egyik a belső, a másik a külső tulajdonságok. Az előbbi kategóriába olyan elemek tartoznak, mint a szín, alak, forma, kiterjedés, a külső tulajdonságok jelölhetik a származási helyet, célt vagy a feltaláló nevét. Petit (2001) szerint a fogalmak differenciálisak, azaz egy fogalmat az különböztet meg a másiktól, hogy legalább egy tulajdonságában eltér tőle. Ennek szemléltetésére vegyük a *számítógép* és a *laptop* párt, amelyekről azt kell eldönteni, hogy két különböző fogalmat jelölnek-e. A számítógép tulajdonságai közé tartozik, hogy képes programok futtatására, bináris adatokkal dolgozik, amelyekkel számításokat végez, és a színe nem meghatározható. A laptop ugyanazon céllal és módszerrel rendelkezik, mint a számítógép, de hordozható, mérete kisebb, és ezáltal gyakran a teljesítménye is alacsonyabb. A két elem között ezért van annyi különbség, amely miatt ezeket két külön terminusként kell kezelni.

3.1.2. A szakterület fogalma

Ahhoz, hogy megtudjuk, hogy a terminus egy adott szakterülethez köthető-e, ahhoz azt is tudni kell, hogy mit takar konkrétan a szakterület, azaz a domén fogalma. Legegyszerűbb venni a tudományág fogalmát, így a biológia, fizika mind-mind önálló szakterületet alkotnának. Azonban Fóris (2005) felhívja a figyelmet arra, hogy a szakterület fogalma nem feltétlenül feleltethető meg a tudományág fogalmának, a tudományok tekintetében uralkodó interdiszciplinaritás miatt. A Fóris (2005) által említett *magas nyomású folyadék kromatográfia* jól tükrözi ezt a jelenséget: ez a főnévi terminus egyaránt köthető a kémia, az orvostudomány vagy a biológia tudományterületéhez, azonban tárgykör tekintetében csak egyhez tartozik, így a szakterület tudományágak halmazainak metszetét takarja, és a metszetekben lévő szakterületrészek azonos fogalmi rendszerrel rendelkeznek.

Felvetődik az a kérdés is, hogy a „szak" előtag a szakterület esetén mit is takar. Pusztay (2008) szerint a fejlett társadalmakban a szakterület a joggal, a politikával, informatikával stb. kapcsolható össze. Azonban a szerző szerint a szakterület nem csak ezekre a tudományokra korlátozhatók le, hiszen a szamojéd nyelvek rénszarvastartással kapcsolatos kifejezései és az eszkimók esetében a 25-30 *hó*ra utaló kifejezés azt sugallja, hogy ezekben a kultúrákban mást is tekinthetünk szakterületnek, illetve szakkifejezéseknek.

3.1.3. A terminus-fogalom közötti bijektív kapcsolat

A klasszikus definíció szerint egy terminus pontosan egy fogalmat nevez meg, és egyetlen fogalmat csak egyetlen terminus jelölhet. Tapasztalatok alapján azonban ez nem mindig van így, pl. Kis (2005) szerint a *directory* szónak már az informatikán belül is több jelentése van: lehet 'könyvtár' vagy az adatbázisok esetében 'olvasásra optimalizált információhalmaz'. Saját példaként említhetjük még a *page* szót, amely szövegszerkesztők esetében 'oldal', honlapszerkesztők vagy az internet esetében 'lap' vagy 'oldal'. Francia példák keresésére az online elérhető terminológiai adatbázis, a *Grand Dictionnaire Terminologique* adhat kiindulási pontot, amelyben például rábukkanhatunk a *dossier* szó által jelölt különböző fogalmakra. Az adminisztratív ügyekben ennek jelentése olyan dokumentumok összessége amelyek „egy adott témával kapcsolatos információkat tartalmaznak, és amelyeket mappába vagy borítékba helyeznek"[6], az építkezés területén „kis szerkezet az egyszerű vagy dupla lejtésű tetőben, az egy szögben találkozó nagyobb

[6] Ensemble de documents qui contiennent des informations relatives à un même sujet, placés dans une chemise, une enveloppe, etc. (saját fordítás)

felületek csatlakozásánál”[7], a számítástechnikában „egy adatrendezési rendszer olyan eleme, amely lehetővé teszi a felhasználó számára a fájlok, dokumentumok és alkalmazások koherens rendezését”[8], az orvostudományban pedig „függőleges, sima és homogén elem, amelyre egy ülő alany feje, válla és a fenék hátsó része támaszkodik”.[9] Ez arra utalhat, hogy a terminusok poliszémek (amelyet viszont a terminusok definíciója kizár), mert egy-egy terminus több fogalmat is jelölhet. A szakirodalomban (pl. Cabré 1999) úgy próbálják mindezt árnyalni, hogy ezekben az esetekben nem is beszélhetnénk poliszémiáról, hiszen az adott szakterületen belül jelentésük egyértelmű, de a szakterület megválasztása kritikus. Ha csak az informatikát vesszük szakterületnek, akkor a *directory* poliszém szó, de ha azon belül vesszük a fájlrendszer és az adatbázisok terminológiáját, akkor észrevehető, hogy az adott terminus az adott szakterületen belül tényleg egy jelentéssel bír. Cabré (1999) ezért azt az álláspontot ismeri el, miszerint a terminusok nem poliszémek, hanem homonimok lehetnek csak, mert adott szűk szakterületen belül nem lehet többértelműség. Ez a nézőpont viszont körkörös definícióhoz vezet, hiszen akkor a köznyelvi szavak sem lehetnének poliszémek, mert elég csak egy szűk területet találnunk, amelyben az adott szó egyértelmű. Így be kell látnunk, hogy hiábavaló a tökéletes bijekció melletti érvelés, ha a valóságban a többértelműség – még ha kisebb mértékben is – de a szaknyelvi terminológiában is előfordul.

3.2. A terminusok további meghatározásai

Petit (2001) szerint a terminusok meghatározásában jelenleg két nézet uralkodik. Az első takarja a terminológiai megközelítést, ami a 3.1. fejezetben tárgyalt klasszikus nézőpont, a másik a nyelvészeti megközelítés, amely ezt bővíti ki. A nyelvészeti megközelítés szerint nem lehet a terminusokat közvetlenül a fogalmakkal összekötni, mert a közvetlen kapcsolat sok mindenre nem adhat magyarázatot. Lérat (1989) ezért úgy vélekedik, hogy a terminusok azért nem köthetők egyből a fogalmakhoz, mert nemcsak denotálják az adott fogalmat, hanem a konnotációkra is lehetőséget adnak. Erre példa a *szemétgyűjtés* terminus, amely a programozás területén a memória felszabadítására utal, de a szóhoz kapcsolódó konnotációk segítségével a terminus könnyen megjegyezhető, és az általa sugallt kép hasonló, mint amit ez a folyamat a programozásban jelent, azaz Kovács (2001)

[7] Petite construction dans le toit à pente simple ou double, à la jonction de surfaces plus grandes qui se rencontrent suivant un angle. (saját fordítás)
[8] Élément d’un système de classement des données qui permet à l’utilisateur de ranger des fichiers, des documents, des logiciels d’application de façon cohérente. (saját fordítás)
[9] Plan vertical, lisse et uniforme, où sont appuyés la tête, les épaules et l’arrière des fesses d’un sujet assis. (saját fordítás)

szóhasználatával élve „környezetfelidéző hatással rendelkezik”. Ezért Lérat (1989) és Depecker (2000) szerint a saussure-i értelemben vett (Saussure 1998) *signifiant* (jelölő) fogalmát be kell emelni a terminusok definíciójába, tehát nem a terminus, hanem a hozzá tartozó *signifiant-signifié* páros köthető a fogalomhoz, és az előbbi összetétel tartozik a nyelvi réteghez, a másik pedig már nem nyelvi réteg. Ezen módosítás segítségével a szóalakhoz tartozó konnotációk a jelölőn keresztül magyarázhatók.

Cabré (2003) szerint a terminus három különböző komponensből épül fel: kognitív, nyelvi és szociokommunikatív komponens. Igaz, ezek a komponensek a köznyelvi szavakra is érvényesek, de a terminusok esetén ezek sokkal korlátozóbbak, ezért a köznyelvi szavakhoz képest a terminusok ebben térnek el igazán. A terminusok kognitív komponense azt takarja, hogy egy adott terminológiai egység jelentése explicit módon rögzített, és ehhez a jelentéshez egyértelmű a megfeleltetése. A terminusok ezen felül egy adott szakterületen belül egy fogalmi struktúrában szerepelnek, ahol minden fogalom legalább egy másik fogalomhoz van kötve. Cabré (1999) úgy véli, hogy a terminusok között egy ilyen hálóban kétféle viszony létezhet: logikai és ontológiai kapcsolat. Két elem akkor áll egymással logikai kapcsolatban, ha jelentésük megegyezik vagy közel azonos, illetve például konjunkció esetén, tehát ha két elem összekötésével egy újabb terminus jön létre. Ontológiai kapcsolat esetében a két elem közelségét tartalmazza a háló, például az alá- és fölérendeltséget.

A Cabré-féle (2003) komponensrendszer második eleme a nyelvi komponens, amely a terminusok nyelvi korlátozásait tartalmazza. Egy terminus lexikai egység, tehát rá is azok a szabályok érvényesek, mint a szavakra, például ugyanolyan módszerekkel (pl. neologizmus) lehet belőlük új szavakat képezni. A terminusok lehetnek egyszerű vagy összetett elemek, és mint összetett elemek belső szintaktikai összetétellel rendelkeznek: ezek az összetételek megegyezhetnek a köznyelvi elemek összetételével. Szintaktikai kategóriájukat illetően lehetnek főnévi, igei, melléknévi vagy határozószói csoportok. Az adott szakterületen belül egy jelentéssel bírnak, és ez a jelentés korlátozott: eseményt, entitást, tulajdonságot vagy kapcsolatot fejezhetnek ki. A terminusok korlátozott jelentéstartalmával Otman (1995) is egyetért, de szerinte a terminusok két fő kategóriába sorolhatók: az egyik a *terminus technicus*ok (*terme technique*) a másik a tudományos terminusok (*terme scientifique*). A tudományos terminusok elméleti fogalmakat írnak le egy speciális szakterületen belül, míg a *terminus technicus*ok eszközöket, folyamatokat, mérhető megfigyeléseket denotálnak. Cabré (2003) ezt azzal is kiegészíti, hogy a

terminusok ettől függetlenül nem biztos, hogy nyelviek, mert szimbólumok vagy más elemek is lehetnek.

A szociokommunikatív komponens azt írja le, hogy egy terminus akkor terminus, ha olyan szövegkörnyezetben szerepel, ahol terminusok lehetnek, tehát a szaknyelvi kommunikációban. Ez a komponens azt is meghatározza, hogy a terminusok nem elsajátított, hanem tanult elemek, amelyek egy adott szakma elsajátítása során épülnek be. A szövegkörnyezet fontosságát illetően Slodzian (2000) azt állítja, hogy egy terminus csak azért terminus, mert szaknyelvi szövegben szerepel. Ezzel egy nem normatív hozzáállást választ, amely a terminusokat előíró szemszögből közelíti meg. Szerinte tévesek azok a megközelítések, amelyek a terminusokat más kritériumok alapján is el szeretnék választani a köznyelvtől, mert a terminusok esetében is létezik poliszémia és variáció.

Sager (1990) azt állítja, hogy a terminusoknak három fő kritériumnak kell megfelelniük: (1) gazdaságosság (*economy*), (2) pontosság (*precision*), (3) megfelelőség (*appropriateness*). A (1) gazdaságosság azt jelenti, hogy a terminusnak nem kell feltétlenül új szónak lennie: a köznyelvben használatos szavak vagy azok kombinációi is könnyen válhatnak terminussá, például a már korábban is említett *szemétgyűjtés*. A (2) pontosság fogalma azt takarja, hogy egy terminus legyen elég pontos, hogy ne legyen kétértelmű, például a *dugaszolóaljzat* összetett szó esetében a két szó külön-külön nem lenne elég a terminus megfelelő leírására. A (3) megfelelőség pedig a két korábbi kritérium konszolidációja: egy terminus legyen eléggé gazdaságos, de pontos is, ezáltal lesz megfelelő is. A *napraforgómag-olaj* egy pontos terminus lehetne, de nem gazdaságos: mivel a napraforgónak csak a magjából nyerhető ki olaj, ezért a *napraforgóolaj* is megfelelő terminus.

3.3. A terminusok összehasonlítása a köznyelvi egységekkel

A terminusokat is, csakúgy mint a szaknyelvet, gyakran definiálják a köznyelvi szavakkal összevetve. Ezt jól tükrözi Sager (2000), Cabré (1999), Lérat (1989), vagy Petit (2001), mert a terminust legjobban azok a különbségek határozzák meg, amelyek köztük és a köznyelvi szavak között vannak. Mivel ezen szerzők eléggé behatóan foglalkoznak ezen eltérésekkel, így a jellemzőknél nem fejtjük ki bővebben, hogy melyik kinek a nevéhez fűződik.

(1) A lexikai egységek a köznyelv részét képezik, míg a terminusok nagyobbrészt a szaknyelvben fordulnak elő. Erre utal Cabré (2003) szociokommunikatív komponense,

illetve Slodzian (2000) terminusdefiníciója is, aki szerint egy terminust csak ez a tényező különbözteti meg a köznyelvi egységektől.[10]

(2) A lexikai egységek rögzülése természetes folyamat, a nyelv elsajátításának része, azaz ez egy tudattalan folyamat. Ezzel szemben a terminusok agyi befogadása tanulási folyamat része, amely a szakma tanulásának folyamatával egybeesik: a szakmában történő fejlődés kihat a szókincs gyarapodására.

(3) Lényeges különbség, hogy egy lexikai egységet nem validál senki: nem szükséges egy intézet hozzájárulása azok használatához. Egy lexikai egységet maga a nyelvi közösség érvényesít: minél többen használják, és minél jobban elfogadják, a nyelvi rögzülésa annál gyorsabb. Ezzel szemben egy terminust csak egy szűkebb réteg hagyhat jóvá: ezek lehetnek tudósok, terminológusok vagy valamilyen nemzetközi intézet. A lényeg az, hogy aki a jóváhagyást végzi, annak ehhez megfelelő hatalommal kell rendelkeznie, tehát egy informatikai terminust biológiai intézet nem érvényesíthet.

(4) A terminusok univerzálisak, hiszen az általuk jelölt fogalmak is univerzálisak. Ez annak is köszönhető, hogy ma már nemzetközi szabványosító intézetek is közreműködnek egy terminus jóváhagyásában. Ezzel szemben egy lexikális egység mindig egy adott nyelvhez köthető, bár léteznek olyan fogalmak, amelyeket számos nyelven elneveznek.

(5) Egy lexikális egység alapvetően poliszém, jelentése a szövegkörnyezettől függ. Ez azt jelenti, hogy egy adott lexikai szó jelentését csak a környezetében szereplő szavak alapján lehet kikövetkeztetni. Ezzel szemben egy terminus jelentése rögzített, a szövegkörnyezetből kiragadva is csak ezt jelentheti. Ehhez azt a kiegészítést kell tenni, hogy egy szóalak több szakterületen is előfordulhat, és a jelentése csak ennek a területnek az ismeretében állapítható meg.

(6) Egy lexikai egységre lehet értékítéletet adni arra vonatkozóan, az milyen szövegstílusban szerepelhet. Például egy szó tartozhat a beszélt nyelvhez, az emelkedett nyelvhasználathoz stb., azonban egy terminusról nem lehet értékítéletet mondani, mert kizárólag a szaknyelvi kommunikációhoz köthető.

3.4. A terminusok definíciói az automatikus terminológiakivonatolásban

A TE-vel kapcsolatos cikkek nagyon keveset foglalkoznak a terminusok meghatározásával: legtöbbször csak azt taglalják, milyen módszereket alkalmaznak a

[10] Ettől függetlenül egy terminus előfordulhat köznyelvi szövegekben is, azonban a definíció szerint akkor elveszti terminusi jellegét.

terminusok automatikus kinyerésére, azzal pedig nem, hogy konkrétan miket is nyernek ki. A terminus minél pontosabb meghatározása azért lenne fontos, mert minden publikáció végén szerepelnek azok az adatok, amelyek egy adott kivonatoló hatékonyságát mérik. Ha a terminusok fogalma nincsen pontosan tisztázva, akkor ez magukat az eredményeket is megkérdőjelezheti. Ezt sokszor úgy próbálják meg áthidalni, hogy a végső ellenőrzésnél úgynevezett bírákat alkalmaznak, akik vagy terminológusok, vagy egy adott szakterület képviselői, és az ő véleményükre alapozva döntik el, hogy egy adott terminusjelölt tényleg az-e. Ekkor tényleg elkerülhető a pontos definíció, mert az ellenőrzési folyamatban ez már nem a szerzők, hanem külső személyek felelőssége. Szintén nem szükséges egy konkrét definíció olyan esetben, amikor az értékelési folyamatban egy külső adatbázis segítségét használják fel. Ekkor az alkalmazás kimenetét összehasonlítják egy terminológiai adatbázissal vagy glosszáriummal: az eredmények itt szintén nem a szerzők terminusról alkotott képétől függnek, hanem a külső referencia pontosságától. Ez a módszer viszont nem teljesen megbízható, mert ha az adatbázis nem elég naprakész, sok olyan elem hiányozhat belőle, amely ténylegesen terminus, de még túl új ahhoz, hogy ezekben szerepeljen. A leggyakoribb valószínűleg azonban mégsem ez a kettő, hanem az, hogy a szerzők saját maguk ellenőrzik le a kimenetet (de ez általában nincs explicit kimondva). Azonban ebben az esetben elengedhetetlen egy olyan meghatározás, amely alapján a későbbi eredményeket megindokolják, mert ellenkező esetben nem megbízható az eredmény.

Azonban a szakirodalom a terminusdefiníciót gyakran mellőzi, és ha van is, akkor is csak a 3.1. alfejezetben kifejtett klasszikus terminus-fogalom kapcsolatot említik meg, azaz egy terminus akkor terminus, ha fogalom tartozik hozzá, és szaknyelvhez köthető (pl. Meilland és Bellot 2005, Foo 2009). Meilland és Bellot (2005) a terminusról azt is állítja, hogy az – a köznyelvi szóval ellentétben – inherens referenciával rendelkezik egy adott szakterülethez. Foo (2009) szerint a terminusok azon elemek halmaza, amelyek egy adott dokumentumot a lehető legjobban meghatároznak. Ez utóbbi definíció a dokumentumindexelés szempontjából előnyös: ha rendelkezünk egy olyan szólistával, amely egy adott dokumentumot meghatároz, akkor a megfelelő dokumentum megkeresése hatékonyabb lesz.

A terminológiakivonatolással kapcsolatosan Jacquemin és Bourrigault (2003) adja a legrészletesebb terminusdefiníciót. Szerintük a klasszikus terminusdefiníció a TE szemszögéből nem alkalmazható, ugyanis a fogalomorientált megközelítés csak a

terminológia normalizálására, illetve szabványosítására használható. Másrészről a probléma az, hogy még most, a mesterséges intelligencia korában sem képes a számítógép megállapítani, mi az a fogalom, amely egy mentális konstrukció, és ezért a terminusok automatikus kinyerését sem lehet erre alapozni. Ezért Jacquemin és Bourrigault (2003) amellett érvelnek, hogy a korpusz alapú terminológiában a terminus a terminológiai elemzés kimenete. Tehát az egyelemű *sejt* vagy a többelemű *vörösvérsejt* azért terminusok, mert azokat korábban manuálisan annak jelölik ki, és ezt a döntést kutatók vagy terminológusok hozzák, így ilyen definícióval nem kell foglalkozni.

Ezenkívül a terminológiakivonatoló alkalmazások céljától függően egyéb terminusdefiníciók is előfordulhatnak. Ha a TE célja a dokumentumindexelés, akkor a terminusok azok az elemek, amelyek egy szöveg tartalmát a legjobban leírják. A dokumentumindexelés folyamán az a fontos, hogy minél több kifejezést gyűjtsünk, amelyek az adott szöveget könnyen elérhetővé teszik a keresés szempontjából (Jacquemin és Bourrigault 2003). Ha a terminológiakivonatoló fordítási munka elősegítésére jön létre, akkor legegyszerűbb, ha azt vesszük terminusnak, amelyeket mindig ugyanúgy kell fordítani (Kis B. 2005), mert a fordítók számára összeállított lista elsősorban azt a célt szolgálja, hogy azokat a kapcsolatokat, amelyeknek egy célnyelvi megfelelője van, az a fordításban is úgy szerepeljen.

3.5. A terminusok jellemzői

Összetettségét tekintve egy terminus lehet összetett vagy egyszerű (L'Homme 2004), egy másik kategorizáció szerint pedig egyszavas vagy többszavas (pl. Boulaknadel és mtsai 2008). Ebben az alfejezetben a terminusok összetettségét, lexikai és szintaktikai jellemzőit mutatjuk be.

3.5.1. A terminusok összetettsége

A terminológiai tanulmányok szerint (pl. Cabré 1999, L'Homme 2004) a terminusok egyszerűségét morfológiai szempontok alapján lehet megállapítani. Egy terminus akkor egyszerű, ha a szótári alakja egy és csak egy morfémát tartalmaz, ellenkező esetben pedig összetett. Tehát a nyelvtani morfémák közül a definíció alapján kizárjuk az inflekciókat, és csak a képzőket vesszük figyelembe. Ebből a szempontból a *body* 'test' terminus egy, az *antibody* 'antitest' már két morfémából tevődik össze, tehát az első egyszerű, a második a derivációs prefixum miatt összetett terminusnak minősül. Cabré (1999) szerint a

terminusok belső morfológiai szerkezete megegyezik a köznyelvi elemekével, amely a következő:

$$(\text{előtag})_1\ \text{szótő}\ (\text{szótő})_{1..n}\ (\text{utótag})_{1..n}$$

Ez azt jelenti, hogy csak a szótő kötelező egy terminus esetében, amelyet megelőzhet legfeljebb egy előtag, és követhet bármennyi másik szótő, amelyet bármennyi utótag követhet. Az összetett terminusokhoz pedig azok a terminusok tartoznak, amelyek több morfémából állnak, de megnyilvánulásuk tekintetében lehetnek egy- (pl. *photosynthesis* 'fotoszintézis') vagy többszavasok (pl. *dialog box* 'párbeszédablak') is.

Az egyszerű és összetett terminusok elkülönítése a TE szempontjából viszont nem ebben a formában létezik. A szakirodalom alapján inkább egyszavas (SWT, *single word term*) és többszavas (MWT, *multi word term*) terminusokat különböztetünk meg, mert a terminusok kinyerésekor ritkán alkalmaznak csak morfológiai elemzőt (pl. Boulaknadel és mtsai 2008). Itt már nem a morfémák száma, hanem a terminus szavainak száma alapján teszünk megkötéseket. Ez azért hasznosabb a terminológiakinyerés szempontjából, mert mint ahogy azt a 5. fejezetben is említeni fogjuk, más módszerek használhatók az egy-, illetve a többszavas terminusok kinyerésére. Mint ismeretes a szó is elég tág fogalom, mert több értelmezése lehetséges: tipográfiai szempontból a szó két szóköz (vagy más szóhatároló egység, pl. vessző, pontosvessző) között helyezkedik el, de olyan összetett szavak is egy szónak tekinthetők, amelyek elemei között van szóköz. Erre példa a francia nyelvben a *machine à café* 'kávégép', amely azért számít ebből a szempontból egy szónak, mert az elemek között erős kohézió van, tehát újabb elemet nem lehet közéjük tenni, így a 'kék kávégép' nem *machine bleue à café*, hanem *machine à café bleue*. A terminuskinyeréssel kapcsolatos szakirodalomban azok az összetett terminusok, amelyek olyan egységekből állnak, amelyek mind rendelkeznek nyelvtani kategóriát jelző szófaji címkékkel. Így az angolban az *operational system* 'operációs rendszer' vagy a *dialog box* 'párbeszédablak' összetett terminusok, mert az egyik egy melléknévi és főnévi címkével rendelkező szóból áll, a másik pedig két darab főnévi annotációval bíró szóból.

Az elkülönítés abból a szempontból lényeges, mert léteznek olyan terminológiakivonatoló alkalmazások, amelyek csak összetett terminusokat nyernek ki egy adott szövegből (pl. Boulaknadel és mtsai 2008), mert az általuk használt módszerek csak többszavas terminusokra alkalmazhatók.

3.5.2. A terminusok morfológiai jellemzői

A terminusok bizonyos része (az idegen eredetűek) bizonyos tudományterületeken (pl. kémia, biológia) olyan specifikus morfológiai tulajdonságokkal rendelkezik, amelyek elősegítik azok felismerését. Ez abból adódik, hogy a terminusok egy nagy része latin vagy görög eredetű, magukkal vonva azon nyelvek morfológiai sajátosságait (Banay 1948). Nybakken (1979) szerint a görög és latin eredetű szavak olyan előnnyel is járnak, hogy azonnal szakkifejezésnek véli azokat az olvasó. Ráadásul, mivel ezek a nyelvek nemzetköziek és univerzálisak, megfelelnek a terminusszabványosítás azon elvének, hogy egy terminus legyen nemzetközi.

A terminusok esetében legtöbbször az idegen eredet affixumok formájában nyilvánul meg, de nagyon gyakori, hogy az egész terminus valamelyik nyelvből származik. Banay (1948) szerint görög eredetű szótő a *nausea* 'hányinger', *ophtalmos* 'szem', *kardia* 'szív', míg latin eredetű szótövek a *cella* 'szoba', *cancer* 'rák', *bacillus* 'pálcika', *pulmo* 'tüdő'. Nybakken (1979) felsorolja az összes olyan prefixumot és szuffixumot, amely az orvostudományban előfordulhat. Az előtagok közül megemlíthető a latin eredetű *trans-* 'keresztül', *circum-* 'körül' vagy a görög eredetű *dys-* 'beteg, rossz' vagy a *cata-* 'lenn'. Az utótagok listája is elég bő, latin eredetű az *-ulum* 'eszköz' vagy *-tia* 'állapot', görög eredetű a *-sia, -sys* 'folyamat, tevékenység'. Nemcsak az orvostudomány, hanem a kémia területén is elég sok olyan terminus található, ahol az előtag latin vagy görög eredete is utal annak terminusi mivoltára, például az olyan görög előtagok, mint *iso-, hydro-* vagy olyan utótagok, mint *-id* kifejezetten ezen a területen használatosak (Anstein és mtsai 2006).

A terminusok másik jellemzője, hogy gyakran idegen eredetűek (Cabré 1999), főleg az angolból veszik át őket. Ez szintén megkönnyíti a terminusok (nem angol nyelvű szövegekből történő) automatikus detektálását, mert az idegennek vélt szavakat nagy valószínűséggel terminusnak lehet venni. Korábbi tanulmányunkban (Nagy 2009a) bemutattuk, hogy a programozás területén nem angol nyelvű szövegekben az angol eredetű szavak mind terminusok, például *plug-in, applet* vagy *LinkedList*.

Azonban egyik tényező sem teljesen meghatározó az automatikus terminológiakivonatolást illetően. Nybakken (1979) megemlíti, hogy az orvosi terminusok nagyrészt latin vagy görög eredetűek, vagy rendelkeznek görög elő- vagy utótaggal, de találhatóak köztük nem idegen eredetű szavak is. De a köznyelvben is rengeteg olyan szó található, amely latin vagy görög eredetű, mégsem szakkifejezés, például a *kata-* előtagú

katasztrófa, így csak erre alapozni nem lehet, de ezek nagyban segítik a terminusok kinyerését.

3.6. A terminus definíciója a saját TE-alkalmazásban

Mint ahogy a 3.4. alfejezetben kifejtettük, hogy a terminus definíciója a TE-alkalmazásokban azért fontos, mert az eredmények ismertetésénél – amikor egy adott kivonatoló hatékonyságát mérjük – fontos tudni azt, hogy mi alapján mondjuk meg azt, hogy egy kinyert terminusjelölt valóban terminus-e vagy sem. Ahogy láttuk, ezt legtöbbször bírák végzik terminológusok és/vagy szakterületi képviselők bevonásával, vagy a program egy adatbázisból kérdezi le, hogy az adott terminusjelölt valóban az-e, vagy nincs meghatározva (vagy csak alig) a terminus fogalma.

A saját TE-alkalmazásban a terminusok meghatározásához a Wüster-féle (1976) klasszikus definíciót alkalmazzuk, azaz (1) a terminus kapcsolódik egy (és csakis egy) fogalomhoz, (2) megnevezi ezt a fogalmat és (3) valamilyen szakterülethez köthető. Habár a 3. fejezetben kifejtettük, hogy a terminusok ezen definíciója egy ideális esetet ír le, ez a meghatározás a korpusz kézi annotáláshoz (terminusok kézi bejelöléséhez) már megfelelő lehet. Ahogy láttuk, a terminusok klasszikus jellemzésének legfőbb problémája az, hogy eszerint a terminus csak egy fogalomhoz tartozik, ami nem mindig igaz. Azonban, ha egy adott terminus egy szakterülethez köthető, az már megfelelő támpont annak kézi annotálásához.

A nem egyértelmű esetekben nagyon hasznos volt még számunkra a *Grand dictionnaire terminologique* online terminológiai szótár, amelyben ezen lehetséges terminusjelölteket le tudtuk ellenőrizni, mivel ez a szótár szinte az összes tudományterületről tartalmaz terminusokat, köztük azon doménekről is, mint az általunk kiválasztott informatika vagy élelmiszeripar.

A jelen fejezetben említett további definíciók, pontosítások legtöbbször csak a terminológia tudományágának azon területeinél lehet fontos, amely a fogalmak és a hozzájuk tartozó terminusok közötti kapcsolat leírásával – ilyen például Lérat (1989) konnotációs definíciója vagy a Cabré (2003)-féle komponensrendszer – vagy a terminusok létrehozásával foglalkoznak: ilyen például a Sager-féle (1990) terminusalkotási kritériumrendszer.

4. A francia nyelvű főnévi terminusok szerkezete

A jelen fejezet áttekinti a francia nyelvű főnévi csoportok és a főnévi terminusok közötti különbségeket a prepozíciós komplementumok (ide tartozik a főnévi fejet követő prepozíció+főnév szerkezet is az összetett szavak esetében) és a melléknévi adjunktumok tekintetében.

Mivel a saját terminológiakivonatoló szabály alapon nyeri ki a terminusjelölteket, minél pontosabban meg kell határozni a terminusok különböző lehetséges szerkezeteit, és hogy ezek miben különböznek a hagyományos főnévi csoportoktól. Mivel a terminus egy fogalmat jelöl, ezért egy lexikai egységnek tekinthető, amely vagy egyszerű vagy összetett szóként realizálódik. Ezzel szemben a főnévi csoport már egy szintaktikai egység, amely rendelkezik egy kötelező fejjel, amelyhez komplementumok, determinánsok és/vagy adjunktumok kapcsolódnak. Azonban a főnévi csoportok és a főnévi terminusok közötti határ nem húzható meg egyértelműen, ahogy ezt az (1)-ben felsorolt példák is szemléltetik.

(1) a. *une* *affaire* *d'* *État*
ART-INDEF.FEM.SG ügy DE állam
'államügy'

b. *une* *affaire* *de* *l'* *État*
ART-INDEF.FEM.SG ügy DE ART-DEF.SG állam
'az állam ügye'

c. *la* *chaise* *de* *bois* *de* *Marie* *qui* *est*
ART-DEF.FEM.SG szék DE fa DE Marie amely van
semblable *à* *un* *vélo*
hasonló -hoz/-höz ART-INDEF.MASC.SG bicikli
'Marie faszéke, amely egy biciklihez hasonlít'

Ezen három elem közül az (1a) lehet terminus és főnévi csoport is, mert a többi (1b-c) csak főnévi csoport. A terminusok is rendelkezhetnek további elemekkel, mint például melléknevekkel, de csak azon melléknevek lehetnek a terminusok részei, amelyek a főnévi fejek egy alfaját jelölik, mint a *filaire* 'vezetékes' a (2a)-ban. Mivel a (2b) a hálózatok nem egy alfaját jelöli, a *grand* 'nagy' melléknév nem része a terminusnak.

(2) a. *le* *réseau* *filaire*
ART-DEF.MASC.SG hálózat vezetékes
'vezetékes hálózat'

b. *le* *grand* *réseau*
ART-DEF.MASC.SG nagy-MASC.SG hálózat
'nagy hálózat'

A terminusok és a főnévi csoportok megkülönböztetéséhez először bemutatjuk a francia főnévi csoportok szerkezetét, különös hangsúlyt fektetve a főnévi fejhez kapcsolódó

prepozíciós komplementumokra és a melléknévi csoportokra, mert főleg ezek azok az elemek, amelyek esetében a főnévi csoportok és a terminusok eltérnek egymástól. Ezen két konstituens bemutatása után ugyanezt a leírást elvégezzük a főnévi terminusokra is.

Alapfeltételezésünk szerint (és ahogy azt (1-2) is mutatja) a főnévi terminusok egyben főnévi csoportok is, de a köznyelvben is létező főnévi csoportok szerkezetüket tekintve nem mindig esnek egybe a terminusokkal, mert összetételük nagyobb változatosságot mutat. Így a jelen fejezetben az összehasonlítás során arra koncentrálunk, hogy mik azok a szerkezetek, amik nem jellemzők terminusokra (de esetleg a köznyelvi terminusokra igen), hogy a szabály alapú terminuskinyerés során ne alkalmazzunk olyan mintákat, amelyek csak a köznyelvi főnévi csoportokra jellemzők.

Ahogy azt a 4.2.3. fejezetben is megemlítjük, a terminust bevezető determinánsok nem képezik az előbbiek részét, így a főnévi terminusok az N'-kategóriának felelnek meg, ezért igazából a jelen fejezetben a főnévi terminusokat a köznyelvi N' projekciójú csoportokkal vetjük össze.

4.1. A francia főnévi csoportok szerkezete

A jelen alfejezet Riegel és mtsai (2009) alapján mutatja be a francia főnévi csoportok szerkezetét. A főnévi csoport kötelező eleme a főnévi fej, de ezen kívül tartalmazhat még adjunktumokat, komplementumokat, valamint specifikálókat. A specifikálók determinánsok, amelyek a főnévi fejet megelőzik, és amelyekből legfeljebb kettő lehet meghatározott sorrendben, mint ahogy azt a (3)-ban lévő példasorozat szemlélteti.

(3) a. *les enfants* ’a gyerekek’
b. *deux enfants* ’két gyerek’
c. *les deux enfants* ’mindkét/a két gyerek’
d. *tous les enfants* ’minden gyerek’
e. **tous les deux enfants*
f. **les tous enfants*

A francia nyelvben a determinánsok használata köznevek esetében szinte kötelező (4a-b), habár bizonyos esetekben és ritkán elhagyható (4c). A tulajdonnevek esetében pedig bizonyos típusú főneveknél, mint országnevek vagy folyónevek (4d), kötelezőek, a többi típusnál nem jellemző (4e)[11].

(4) a. *Je mange du fromage.*
én eszem ART-PART.MASC.SG sajt
’Sajtot eszem.’

[11] A disszertációban nem térünk ki a pontos szabályokra azok irreleváns jellege miatt.

b. *Il voit <u>des</u> voitures.*
ő-MASC lát ART-INDEF.PLUR autók
'Autókat lát.'
c. *Pierre est <u>Ø</u> ingénieur.*
Pierre van mérnök
'Pierre mérnök.'
d. *Il aime <u>la</u> Hongrie.*
ő-MASC szereti ART-DEF.FEM.SG Magyarország
'Szereti Magyarországot.'
e. *Pierre aime (?<u>la</u>) Marie.*
Pierre szereti (?ART-DEF.FEM.SG) Marie
'Péter szereti Marit.'

A komplementumok két nagyobb csoportra oszthatók: prepozíciós frázisok (5a) és alárendelő mondatok, amely utóbbiak infinitívuszi (5b), illetve kompletív, finit igét tartalmazó mellékmondatok (5c) lehetnek.

(5) a. *la tarte <u>aux prunes</u>*
ART-DEF.FEM.SG sütemény À+ART-DEF.PL szilvák
'szilvás sütemény'
b. *l' idée <u>de vérifier</u>*
ART-DEF.SG ötlet DE ellenőrízni
'az ellenőrzés ötlete'
c. *le fait <u>qu' il soit parti</u>*
ART-DEF.MASC.SG tény hogy ő-MASC van-PRES.SUBJ elmegy-PAST.PART.MASC
'az a tény, hogy elment'

A komplementumok helye és száma meghatározott: mindig a főnévi fej után állnak, és általában egy lehet belőlük.

Az adjunktumok lehetnek melléknévi csoportok (6a), vonatkozói mellékmondatok (6b) és prepozíciós szintagmák (6c).

(6) a. *un film <u>intéressant</u> / un <u>intéressant</u> film*
ART-INDEF.MASC.SG film érdekes-MASC.SG
'egy érdekes film'
b. *la personne <u>que nous avons vue</u>*
ART-DEF.FEM.SG személy akit/amit mi AUX lát-PAST.PART.FEM
'a személy, akit láttunk'
c. *la personne <u>dans l' armoire</u>*
ART-DEF.FEM.SG személy -ban/-ben ART-DEF.SG szekrény
'a szekrényben lévő személy'

Az adjunktumok helye és száma már nem annyira kötött, mint a komplementumok esetében. A prepozíciós szerkezetek és a vonatkozói mellékmondatok csak a főnévi fej után állhatnak, míg a melléknevek a főnév előtt és után is. Az adjunktumok száma egy főnévi csoporton belül nem korlátozott.

Mivel a terminológiakivonatolásnál fontos a szintaktikai összetétel minél pontosabb feltérképezése, ezért a két legfontosabb főnévicsoport-összetevőt, a prepozíciós szintagmákat és a melléknévi csoportokat, a következőekben bővebben is kifejtjük.

4.1.1. Prepozíciós szintagmából álló komplementumok

A prepozíciós szintagmák meghatározása a francia nyelv esetében nem evidens, mert nem mindig lehet egyértelműen megkülönböztetni a főnévi csoporton belül új entitást bevezető prepozíciós szintagmát (7a) és a főnévi fejhez kapcsolódó, és azzal egy összetett főnevet alkotó prepozíciós szintagmát (7b). Az összetett főnevek esetében (7b) a prepozíciót egy determináns nélküli főnév követi, mivel a determináns megléte egy összetett főnévi csoporthoz vezetne, ahol a prepozíciót követő főnévi csoportot egy önálló referenciával rendelkező, beágyazott főnévi csoportnak tekinthetjük (7a). Azonban az összetett főnevek prepozíciós csoportjai nem referenciálisak. A determinánsnak tehát alapvető szerepe van abban, hogy megkülönböztesse a főnév+prepozíció+főnév szerkezetű összetett főneveket (7b) a prepozíciós szintagmát tartalmazó főnévi csoportoktól (7a).

(7) a. *le* *moulin* *du* *village*
ART-DEF.MASC.SG malom DE+ART-DEF.MASC.SG falu
'a falu malma'

b. *le* *moulin* *à* *vent*
ART-DEF.MASC.SG malom À szél
'szélmalom'

Riegel és mtsai (2009) a prepozíciós csoportok közé minden olyan szintagmát besorolnak, amelyek esetében a prepozíciót egy egész főnévi csoport követ (pl. *le chien* $_{PP}$[*de* $_{NP}$[*la voisine*]] 'a szomszéd kutyája'). Azonban a példái között olyan összetett szavakat is találunk, mint *canne à pêche* 'horgászbot', ahol a *pêche* 'horgászás' nem alkot önállóan főnévi csoportot. Azonban bizonyos nézőpontok szerint (pl. Bosredon és Tamba 1991) szemantikai szempontból az összetett főnevek egyszerű főnevek, de formálisan főnévi csoportok. Ezért Bosredon és Tamba (1991) megkülönbözteti a hagyományos prepozíciós szintagmákat a főnévhez kapcsolódó prepozíció+főnév szekvenciáktól: az előbbieket konstituensnek az utóbbiakat formánsoknak nevezi.

A továbbiakban prepozíciós szintagmákon mind a prepozíciós konstituenseket, mind a formánsokat értjük, mert formáns és konstituens között a formájukat tekintve nem lehet egyértelmű határt húzni. Egyrészről vannak olyan prepozíciók, amelyeket általában (vagy bizonyos esetekben) nem követ determináns (pl. *sans* 'nélkül' vagy *en* '-ba(n)/-be(n)'), de nem alkotnak az előttük lévő főnévvel összetett főnevet (8a). Másrészről

vannak olyan összetett főnevek, amelyek olyan prepozíciós szintagmát tartalmaznak, amelyekben a prepozíciót determináns követi (9a).

(8) a. *voyage en Italie*
utazás -ba(n)/-be(n) Olaszország
'olaszországi utazás'
b. **voyage en l' Italie*
utazás -ba(n)/-be(n) ART-DEF.SG Olaszország

(9) a. *cancer de la peau*
rák DE ART-DEF.FEM.SG bőr
'bőrrák'
b. ?*cancer de peau*

A jelen fejezet további részében a determináns nélküli főnévi csoportokra koncentrálunk, melyek nagyobb eséllyel lehetnek terminusok.

A francia nyelv egyik tulajdonsága, hogy a szóösszetételeket a legtöbb esetben prepozíciós szerkezettel oldja meg a szavak közvetlenül egymás után történő helyezésével ellentétben. Az összetett főnevek nem első tagjai általában főnevek (10a-d), de lehetnek főnévi igenevek is (10e), amelyeket a főnévi fejhez prepozícióval kötünk. Az összetett főnevek tagjait általában a *de* prepozíció köti össze (10a-b), de bizonyos esetekben a két elem között más prepozíció is lehet (pl. az *en* a (10c)-ben vagy az *à* a (10d-e)-ben). Az összetett főnevek írásakor nem használunk kötőjelet, kivéve néhány esetben, mint a (10c)-ben (Riegel és mtsai 2009).

(10) a. *hôtel de ville*
ház/hotel DE város
'városháza'
b. *professeur de français*
tanár DE francia
'franciatanár'
c. *arc- en- ciel*
ív -ba(n)/-be(n) ég
'szivárvány'
d. *rouleau à pâtisserie*
henger À sütemény
'sodrófa'
e. *machine à laver*
gép À mosni
'mosógép'

A kötőjel jelenléte nem csak pusztán helyesírási kérdés. A terminológiakivonatoló automatikus annotációkkal dolgozik, és ezek az annotáló programok (mint például az is, amelyet mi használunk) nem bontja szét a kötőjellel írt szóösszetételeket külön egységekre, hanem egy szónak tekinti őket, amelyeket főnév címkével lát el. Így a (10c) típusú összetett szavak kinyeréséhez elég az egyszavas terminusok (pl. *réseau* 'hálózat')

kinyeréséhez használt minta.

A francia nyelvben is léteznek azonban prepozíció nélküli szóösszetételek, melyek írásmódja történhet kötőjel nélkül (11c), illetve kötőjellel is (11a-b).

(11) a. *le* *gratte-ciel*
ART-DEF.MASC.SG kapar-ég
'felhőkarcoló'
b. *le* *chou-fleur*
ART-DEF.MASC.SG káposzta-virág
'karfiol'
c. *la* *pause* *déjeuner*
ART-DEF.FEM.SG szünet ebéd
'ebédszünet'

4.1.2. Melléknevek helye a főnévi csoportban

A melléknevek a francia nyelvben állhatnak a főnév előtt és a főnév után is. Cinque (1994) úgy fogalmaz, hogy az újlatin nyelvek, és ezáltal a francia is, az ANA típusú nyelvek közé tartozik, míg a germán nyelvek inkább AN típusúak, tehát az utóbbi nyelvekben a melléknév prenominális, az előbbi nyelvekben pedig lehetnek pre- és posztnominálisak is. Ezt szemléltetik a (12)-ben szereplő főnévi csoportok: a francia nyelvben a melléknév állhat főnév előtt vagy után (12a), de az angolban csak a főnév előtt (12b-c).

(12) a. *la* *jolie* *chambre* *bleue*
ART-DEF.FEM.SG szép-FEM.SG szoba kék-FEM.SG
'a szép kék szoba'
b. *the nice blue room*
c. *the nice room blue

Jóllehet, a melléknevek a franciában kerülhetnek a főnév elé és után is, azonban vannak rájuk nézve megszorítások. Alapértelmezés szerint a melléknév a főnév után található (13a), de előre is kerülhet, ha hangsúlyos, vagy más szóval fokalizált (13b) (Laenzlinger 2003).

(13) a. *un* *roman* *ennuyeux*
ART-INDEF.MASC.SG regény unalmas-MASC.SG
'egy unalmas regény'
b. *un ennuyeux roman*

Bizonyos esetekben a pre- és posztnominális mellékneveket másképpen értelmezzük. Bouchard (1998) szerint a főnév után álló melléknevek a főnévre mint egészre vonatkoznak, míg ugyanaz a melléknév prenominálisként a főnév specifikus belső jelentés-összetevőit módosítja. (14) és (15) azokat a tipikus eseteket szemlélteti, amelyeket a franciát mint idegen nyelvet tanulóknak szokás klasszikus példaként felhozni:

(14) a. *mon* *fauteuil* *ancien*
az én ...-m fotel régi
'a régi fotelem'
b. *mon ancien fauteuil*
'egy [pl. nekem] régi fotel'

(15) a. *un* *parent* *seul*
ART-INDEF.MASC.SG szülő egyetlen
'egy egyedül lévő szülő'
b. *le seul parent*
'az egyetlen szülő'

Ezen példák tisztán mutatják a melléknevek pre- és posztnominális változatai közötti jelentésbeli különbséget. Például (14a) 'egy valóban régi fotelt' jelent, míg (14b) egy 'olyan fotelt, amely rég jelent meg egy meghatározott környezetben, de nem feltétlenül régen gyártott'. (15a) egy olyan szülőt jelöl, aki egyedül van/él, míg (15b) egy olyan szülőt, aki egy adott környezetben van egyedül, például egy csoportban.

Néhány melléknév mindig megelőzi a főnevet: ezek általában "rövidek", egy- vagy legfeljebb kétszótagú szavak. Ebben az esetben a nyelvtankönyvek fonoritmikus és használatot érintő faktorokkal magyarázzák azok előre történő helyezését, ugyanis ezen mellékneveket gyakran használjuk a mindennapos társalgásban (Laenzlinger 2003). Ilyen melléknevek például a *petit* 'kicsi', *beau* 'szép' vagy a *long* 'hosszú' (16a-c).

(16) a. *une* *petite* *chose*
ART-INDEF.MASC.SG kis-FEM.SG dolog
'egy kis dolog'
b. *une* *belle* *chanson*
ART-INDEF.FEM szép-FEM.SG dal
'egy szép dal'
c. *une* *petite* *belle* *tour*
ART-INDEF.FEM.SG kis-FEM.SG szép-FEM.SG torony
'egy kis szép torony'

A (16c) jelű példa még azt is igazolja, hogy ezen melléknevek közül többet is lehet egyszerre előre helyezni. Azonban, ha ezeket a mellékneveket valamilyen prepozíciós komplemens követi, kötelezően a főnév mögé kerülnek (17).

(17) a. *une* AP[*longue*] *rivière*
ART-INDEF.FEM.SG hosszú-FEM.SG folyó
'egy hosszú folyó'
b. *une* *rivière* AP[*longue* *de* *300* *mètres*]
ART-INDEF.FEM.SG folyó hosszú-FEM.SG DE 300 méter-PL
'egy 300 méter hosszú folyó'
c. *une* *rivière* AP[*plus* *longue* *que* *le* *Nil*]
ART-INDEF.FEM.SG folyó -bb hosszú-FEM.SG mint ART-DEF.MASC.SG Nílus
'egy Nílusnál hosszabb folyó'
d. **une* AP[*longue de 300 mètres*] *rivière*

Egy prenominális melléknév akkor is posztnominális lesz, ha határozószó módosítja azt. Ez abban az esetben nem igaz, ha mind a melléknév, mind a határozószó rövid és gyakran használt (19). Ilyen határozószó lehet a *tout* 'teljesen', *très* 'nagyon' vagy *trop* 'túl', amelyek esetében az AP helye a főnévi csoporton belül fakultatív (18).

(18) a. *une* *courte* *enfance*
ART-INDEF.FEM.SG rövid-FEM.SG gyerekkor
'egy rövid gyerekkor'
b. *une* *très* *courte* *enfance* / *une enfance très courte*
ART-INDEF.FEM.SG nagyon rövid-FEM.SG gyerekkor
'egy nagyon rövid gyerekkor'

(19) a. *une* *enfance* *extrêmement* *courte*
ART-INDEF.FEM.SG gyerekkor túlságosan rövid-FEM.SG
'egy túlságosan rövid gyerekkor'
b. **une* AP[*extrêmement courte*] *enfance*

Bouchard (1998) szerint a képzett melléknevek, úgymint a melléknévi szerepben lévő jelen vagy múlt idejű melléknévi igenevek, is csak főnév után állhatnak (20a-b).

(20) a. *un* *morphème* *lié*
ART-INDEF.MASC.SG morféma köt-PART.PAST.MASC.SG
'kötött morféma'
b. *une* *chaise* *roulante*
ART-INDEF.FEM.SG szék gurul-PART.PRES.FEM
'kerekesszék'

A képzett intenzionális melléknevek (amelyek eredetileg igenevek) látszólag ellentmondanak az előbbi szabálynak, mert ezek a francia nyelvben csak prenominálisok lehetnek. Ezt a (21)-ban szereplő példák is igazolják.

(21) a. *un* *prétendu* *chef* *d'* *Etat*
ART-INDEF.MASC.SG állít-PART.PAST.MASC.SG főnök DE állam
'egy állítólagos államfő'
b. *un* *soi-* *disant* *dentiste*
ART-INDEF.MASC.SG PRON-REFL.3SG.GENER mond-PART.PRES.MASC.SG fogorvos
'egy egyéni stílusú fogorvos'

Riegel és mtsai (2009) szerint a melléknevek főnévhez viszonyított helye a francia nyelvben a melléknév jelentésétől is függhet. Négy típusú melléknevet különböztetünk meg: csoportosító, nem csoportosító, relációs melléknevek és sorszámnevek[12]. A csoportosító melléknevek olyan tárgyilagos tulajdonságokat jelölnek, amelyek alapján a főnevet egy adott kategóriába lehet helyezni: ilyenek a színt, formát, alakot jelentő szavak, mint a *vert* 'zöld' vagy *ovale* 'ovális'. A nem csoportosító melléknevek egy szubjektív tulajdonságot jelölnek, például *petit* 'kicsi', *intéressant* 'érdekes'. A relációs melléknevek főnévből képzettek, és jelentésüket tekintve azonosak a *de*+főnév szerkezettel, ilyen

[12] *Adjectifs classifiants, adjectifs non-classifiants, adjectifs relatifs, adjectifs ordinaux.*

például a *régió*n ’régió’ szóból képzett *régional* ’regionális’, amely a *langue régionale* ’regionális nyelv’ esetében ekvivalens a *langue de région* ’a régió nyelve’ alakkal. A francia nyelvtanok szerint a sorszámnevek szintén melléknevek, ilyenek a *premier* ’első’ vagy a *deuxième* ’második’.

A fenti négy melléknévtípus közül a csoportosító (22a) és a relációs (22b) melléknevek (amelyek a bouchard-i (1998) terminológia szerint interszektív predikatív melléknevek) csak a főnevek után állhatnak, a nem csoportosító melléknevek (22c) és a sorszámnevek (22d) helye a főnévhez képest fakultatív (de az utóbbiak esetében az alapértelmezett a főnév előtti hely).

(22) a. *un* *bureau* *ovale* / **un ovale bureau*
ART-INDEF.MASC.SG iroda ovális
’ovális iroda’
b. *une* *langue* *régionale* / **une régionale langue*
ART-INDEF.FEM.SG nyelv regionális-FEM.SG
’egy regionális nyelv’
c. *un* *livre* *intéressant* / *un intéressant livre*
ART-INDEF.MASC.SG könyv érdekes-MASC.SG
’egy érdekes könyv’
d. *la* *langue* *première* / *la première langue*
ART-DEF.FEM.SG nyelv első-FEM.SG
’az első nyelv’

Hasznos még megvizsgálni a melléknévi adjunktumok és a prepozíciós szintagmák egymáshoz kötődő disztribúcióját a főnévi csoporton belül azért, hogy minél pontosabban feltérképezzük a lehetséges főnévicsoport-mintákat: a főnévi terminusok kinyerésénél ugyanis konkrét mintákat fogunk alkalmazni, amelyeknél fontos ezen elemek sorrendjének minél pontosabb meghatározása. A prepozíciós szerkezetek ebből a szempontból nem problémásak, mert a főnévi fej után állnak. A kérdés csak az, hogy a posztnominális melléknevek ezeket a prepozíciós frázisokat megelőzik-e vagy sem. Laenzlinger (2003) szerint egy melléknév kerülhet a főnévi fej és a prepozíciós szintagma közé, de a komplementum mögé is, így a helye fakultatív. Ezt szemlélteti (23):

(23) a. *un* *ministre* *de la* *Justice* *blanc*
ART-INDEF.MASC.SG miniszter DE ART-DEF.FEM.SG igazságügy fehér-MASC.SG
’egy fehér igazságügyi miniszer’
b. un ministre blanc de la Justice

Ugyanakkor, ha a főnévi fej és a prepozíciós komplementum együttesen egy lexikálisan rögzített entitást fejeznek ki, akkor a melléknév nem kerül közéjük, ahogy ezt a (24)-ben lévő példa is mutatja, ahol a *lunettes de soleil* ’napszemüveg’ az összetartozó lexikális egység.

(24) a. *des* *lunettes* *de* *soleil* *nouvelles*
ART-INDEF.PL szemüveg-PL DE nap új-FEM.PL
'új napszemüveg'
b. ??*des lunettes nouvelles de soleil*

Abeillé és Godard (1999) egy másik megközelítésben vizsgálja a főnévi fejek és azok melléknévi adjunktumainak relatív disztribúcióját a francia főnévi csoportokban. Bevezetik a relatív súly (*relative weight*) fogalmát, aminek segítségével már számot tudnak adni a melléknevek főnévi fejhez viszonyított helyéről. A melléknevek helyét a főnévi fejhez viszonyítva azon melléknevek és/vagy a főnévi fejek súlya határozza meg. Kétféle típusú súly létezik: „lite" (könnyű) és „non-lite" (nehéz). Azt, hogy egy melléknév könnyű vagy nem[13], azt (1) vagy a lexikon határozza meg (bizonyos melléknevek könnyűek, bizonyosak nem, és számos melléknévnél ez a tulajdonság nem meghatározott), (2) vagy azon szabályok, amelyek a frázisok szintaktikai szerkezetét írják le (a legtöbb frázis nehéz, mások könnyűek). Például létezik egy szabály, ami leírja, hogy a könnyű melléknevek kötelezően prenominálisak, a nehezek pedig posztnominálisak.

(25) a. *une* lite A[*belle*] *dame*
ART-INDEF.FEM.SG szép-FEM.SG hölgy
'egy szép hölgy'
b. *une* *femme* non-lite A[*russe*]
ART-INDEF.FEM.SG hölgy orosz
'egy orosz hölgy'

Azonban könnyű melléknevek nem köthetők nehéz főnevekhez (mint például mellérendelt főnevekhez): ezen esetekben csak nehezek lehetnek, tehát kizárólag a főnév után állhatnak, ahogy ezt a (26)-ben szereplő példa is mutatja:

(26) non-lite N[*des* *hommes* *et* *des* *enfants*] non-lite A[*jolis*]
ART-INDEF.PL férfiak és ART-INDEF.PL gyerekek kedves-MASC-PL
'kedves férfiak és gyerekek'

A koordinált melléknevek relatív súlya nem mindig teljesen egyértelmű. Két koordinált könnyű melléknév súlya nem meghatározott, tehát meg is előzheti és követheti is a főnevet (ahogy ezt (27) is mutatja), és a lexikonban nem meghatározott súlyú koordinált melléknevek rendszerint nehezekké válnak, amit a (28)-es példa szemléltet.

(27) a. *une* A non.det.[light-A[*jolie*] *et* light-A[*belle*]] *chambre*
ART-INDEF.FEM.SG kedves-FEM.SG és szép-FEM.SG szoba
'egy szép és kedves szoba'
b. *une chambre* A non.det.[light-A[*jolie*] *et* light-A[*belle*]]

[13] A melléknevek esetében a relatív súly nem feltétlenül a melléknevek lexikonban tárolt súlyát jelenti: egy lexikailag könnyű melléknév bizonyos esetekben nehézzé válhat (pl. 26), így *könnyű* és *nehéz melléknév* alatt a melléknevek viselkedését is leírhatjuk az adott kontextusban, nemcsak a lexikonban tárolt súlyukat.

(28) a. ??*une* non-light A[A non.det.[*excellente*] *et* A non.det.[*joyeuse*]] *femme*
ART-INDEF.FEM.SG kiváló-FEM.SG és boldog-FEM.SG nő
'egy kiváló és boldog nő'
b. *une femme* non-light A[A non.det.[*excellente*] *et* A non.det.[*joyeuse*]]

4.2. A francia főnévi terminusok belső szerkezete

Az 4.1. fejezetben ismertettük a francia főnévi csoportok belső szerkezetét, különös hangsúlyt fektetve az azokban előforduló melléknévi, illetve prepozíciós szerkezetekre. E gondolatmenet mentén mutatjuk be minél pontosabban a főnévi terminusok szerkezetét.

Mivel a francia nyelvben nincs olyan megkülönböztető jegy, ami egyértelműen elkülönítené a terminusokat a köznyelvi főnévi csoportoktól, ezért a minta alapú terminuskinyeréskor szükség van a főnévicsoport-minták szűkítésére, aminek segítségével nagyobb hatékonysággal nyerhetünk ki főnévi terminusokat. Alkalmazhatnánk ugyan azon általános mintákat is, amelyek főnévi csoportok kinyerésére alkalmasak, azonban a nagyobb pontosság érdekében figyelembe kell venni a következő megfontolásokat.

4.2.1. Prepozíciós szintagmák a főnévi terminusokban

A francia nyelvben nem jellemző a szóösszetételek használata, így a lexikalizálódott prepozíciós komplementumokat is figyelembe kell vennünk. Mint ahogy azt az 4.1.1. fejezetben leírtuk, a lexikalizálódott komplementumokat onnan ismerjük fel, hogy azok rendszerint nem tartalmaznak névelőt, és a prepozíció után főnév vagy főnévi igenév található. Ezekre a (29)-ben adunk példákat:

(29) a. *farine de blé*
liszt DE búza
'búzaliszt'
b. *machine à laver*
gép À mosni
'mosógép'

Ezen elemeket a terminuskivonatolás során mindenképpen egy egységnek kell tekintenünk, hiszen a *farine* és a *blé* külön-külön lehetnek ugyan terminusok, de ha így együtt szerepelnek, ezek egy fogalmat jelölnek. Ha azonban a prepozíció után determináns következik, akkor ott már kétséges a helyzet. Alapértelmezés szerint azokat nem érdemes egységként terminusnak venni, mert ott valószínűsíthető, hogy két külön terminus összekapcsolásáról van szó.

(30) a. *mise à jour du* (de+le) *site web*
tétel À nap DE+ART-DEF.MASC.SG weboldal
'a weboldal[nak a] frissítése'

b. *création* *d' <u>un</u>* *site web*
létrehozás DE ART-INDEF.MASC.SG weboldal
'egy weboldal létrehozása'

Ezekben az esetekben tehát terminusjelöltként kell felvenni a *mise à jour,* illetve a *site web* szókapcsolatokat, ugyanígy a (30b) esetében is. A determináns ezért egyfajta határolóként is szerepelhet, ami különböző terminusjelölteket választ el egymástól.

Cadiot (1993) is ezt támasztja alá: a determinánst nem tartalmazó prepozíciós szintagma az előtte álló főnévnek egy alfaját adja, míg a determinánst tartalmazó változat annak csak egy előfordulását írja le. Az utóbbi esetben a kategorizálás csak közvetett lehet: ez az extenzionális tulajdonságból következik, míg az előbbi esetben intenzionális tulajdonságok alapján közvetlenül tudja azt a főnevet kategorizálni.

(31) a. *chat* *à* *poils* *longs*
macska À szőr-PL hosszú-MASC.PL
'hosszúszőrű macska'
b. *chat* *aux* *poils* *mouillés*
macska À+ART-DEF.PL szőr-PL nedves-MASC.PL
'nedves szőrű macska'

A (31)-ben látható példák jól mutatják, hogy a determináns nélküli (31a) a macskák egy alfaját csoportosítja, míg a determinánssal rendelkező (31b) egy olyan macskát ír le, amely rendelkezik azzal a tulajdonsággal, hogy a szőre nedves, de emiatt ő nem egy új alfaj a macskákon belül.

Hasonlóképpen vélekedik Anscombre (1990, 1991), aki szerint a determináns nélküli prepozíciós utómódosító a főnévi fej egy nélkülözhetetlen tulajdonságát (*propriété essentielle*) írja le, míg a determinánssal rendelkező utómódosító annak egy véletlenszerű tulajdonságát (*propriété accidentielle*). Úgy fogalmaz, hogy a P tulajdonság egy nélkülözhetetlen tulajdonsága az E entitásnak, ha P-t E elválaszthatatlan részének tekintjük. Ezzel ellentétben P véletlenszerű tulajdonság, ha a P-nek egy ideiglenes tulajdonsága. Tehát a nélkülözhetetlen tulajdonság ténylegesen tulajdonság, míg a véletlenszerű tulajdonság csak egy aktuális állapot. A (32)-ben szereplő példák szemléltetik, hogy a *bateau à voiles* 'vitorlás hajó' determináns nélküli prepozíciós utómódosítóval rendelkező összetett főnév után csak bizonyos típusú mellékneveket használhatunk: ezen melléknevek a vitorla típusára kell, hogy vonatkozzanak, míg a determinánssal rendelkező (32d) esetében ideiglenes tulajdonságot kifejező mellékneveknek.

(32) a. *bateau à voiles*
hajó À vitorlák
'vitorlás hajó'
b. *bateau à voiles carrées / latines*
hajó À vitorlák négyszögletű-FEM.PL latin-FEM.PL
'négyszögletű- vagy latinvitorlás-hajó'
c. *bateau à voiles ??hissées / *déchirées*
hajó À vitorlák felvont-FEM.PL elszakadt-FEM.PL
'elszakadt- vagy felvontvitorlás-hajó'
d. *bateau aux voiles hissées / déchirées*
hajó À+ART-DEF.PL vitorlák felvont-FEM.PL elszakadt-FEM.PL
'felvont/elszakadt vitorlás hajó'

(Anscombre 1991: 26)

A példák alapján Anscombre (1991) azt állítja, hogy a mellékneveknek „generikusoknak" kell lenniük, így a határozott névelő megjelenése a vitorlák mint különálló entitások feltételezett létezését vonja maga után.

Cadiot (1993) azt is leírja, hogy hasonló a helyzet, ha nem tartalmaz melléknevet a prepozíciós frázis. A determináns jelenléte azt sugallja, hogy a prepozíció előtt és után álló rész egy külön entitást jelöl, amelyek saját referenciával rendelkeznek. Ezt igazolja az alábbi két pár is:

(33) a. *un bagage à main*
ART-INDEF.MASC.SG poggyász À kéz
'egy kézipoggyász'
b. *un bagage à la main*
ART-INDEF.MASC.SG poggyász À ART-DEF.FEM.SG kéz
'egy poggyász a kézben'

(34) a. *Jean a un bagage à main*
Jean birtokol-3SG ART-INDEF.MASC.SG poggyász À kéz
cependant il le porte au ventre.
de ő-MASC azt-MASC hord-3SG À+ART-DEF.MASC.SG has
'Jeannak van egy kézipoggyásza, de a hasán hordja.'
b. **Jean a un bagage à la main*
Jean birtokol-3SG ART-INDEF.MASC.SG poggyász À ART-DEF.FEM.SG kéz
cependant il le porte au ventre.
'Jeannak van egy poggyász a kezében, de a hasán hordja.'

Anscombre (1991) ezzel kapcsolatban azt állítja még, hogy a determináns nélküli prepozíciós utómódosítókban nem szerepelhet bármilyen főnév, csak olyan, amely a főnévi fej egy nem evidens, belső tulajdonságát írja le. Erre példa a (35)-ben lévő főnévi egység, mert az autó fogalma eleve feltételezi a kormányt, de ha ez az eszköz hidrogén hajtású, akkor azt már meg lehet említeni mint egy alapvető tulajdonságot.

(35) a. *voiture à *volant*
autó À kormány
'kormányos autó'

b. *voiture à hydrogène*
autó À hidrogén
'hidrogén hajtású autó'

(36) a. **un chat à deux oreilles*
ART-INDEF.MASC.SG macska À két fülek
'kétfülű macska'

b. **un vélo à roues*
ART-INDEF.MASC.SG bicikli À kerekek
'kerekes bicikli'

Hasonló elvek alapján hibásak (36a-b) prepozíciós frázisok: az, hogy egy macska két füllel rendelkezik, vagy az, hogy egy bicikli kerékkel rendelkezik, az ezen entitások egy alapvető tulajdonsága. Ha ezen utómódosítókat kibővítjük vagy módosítjuk, helyes főnévi csoportokat kapunk, hiszen egy mutáns háromfülű macska, vagy a négyszögletű-kerekes bicikli esetében a bővítmények nem evidens tulajdonságokra mutatnak rá.

(37) a. *un chat à trois oreilles*
ART-INDEF.MASC.SG macska À három fülek
'háromfülű macska'

b. *un vélo à roues carrées*
ART-INDEF.MASC.SG bicikli À kerekek négyszögletű-FEM.PL
'négyszögletű-kerekes bicikli'

Cadiot (1993) fenti kijelentése a determinánsok külön entitásokat bevezető szerepéről nem mindig állja meg a helyét, ugyanis ritkább esetekben a terminusjelölt tartalmazhat determinánst is azon tagja előtt, amely nem külön entitás. Ezek száma elég csekély, és megjelenésükre sem adható magyarázat. Ezeket a (38)-ben szemléltetjük.

(38) a. *cancer de la peau*
rák DE ART-DEF.FEM.SG bőr
'bőrrák'

b. *vidéo à la demande*
videó À ART-DEF.FEM.SG kérés
'VOD' vagy 'Video on demand'[14]

Nem igazolható, hogy a bőrrák miért nem csak *cancer de peau* a használt *cancer de la peau* helyett. Egy korábbi tanulmányunkban (Nagy 2009c) bemutattuk, hogy a terminusok esetén a belső determinánsok jelenléte általában 7% az összes terminus között, azonban arra vonatkozólag, hogy ezen 7%-nyi terminus között hány szerepelhet determináns nélkül is, nincs empirikus eredményünk. Ebből adódóan, egy kisebb veszteséggel számolva, nem vettük figyelembe a determinánssal rendelkező terminusokat.

A terminusok esetében azt is figyelembe kell venni, hogy melyek azok a prepozíciók, amelyek ezekben potenciálisan előfordulhatnak. Riegel és mtsai (2009)

[14] A kifejezés magyarosított változatai, például 'videó kérésre' vagy 'videó igény szerint' nem terjedtek el.

szerint a franciában előforduló prepozíciók közül a *de* és az *à*, illetve azok összevont alakjai (*de+le* → *du*, *à+les* → *aux*, *de+lequels* → *desquels* stb.) a leggyakoribbak. A korpuszon végzett vizsgálat szerint két olyan prepozíció van – a *sans* 'nélkül' és a *pour* '-ért, vmilyen célból' –, amelynek prepozíciós projekciója (PP) vagy nem a főnévi fejhez kapcsolódik, vagy a főnévi fejhez kapcsolódik, de annak nem jelöli alfaját, azaz annak csak egy szabad bővítménye. Így ezeket nem célszerű felvenni a terminusban előforduló lehetséges prepozíciók közé. A (39a) mutatja a *sans* terminusi és a (39b) a nemterminusi használatát:

(39) a. *effectuer* *un* *diagnostic*
végrehajtani ART-INDEF.MASC.SG diagnosztika
sans *perturber* *le* *réseau*
nélkül zavarni ART-DEF.MASC.SG hálózat
'diagnosztikát végrehajtani a hálózat zavarása nélkül'

b. *police de charactères* ***sans*** *empattement*
betűtípus nélkül talp
'talpatlan betűtípus'

A (39) példa egyértelműen mutatja, hogy az első esetben a *sans* prepozíció utáni rész az *effectuer* ige bővítménye, míg a második esetben a terminus része, de mindkettő prepozíciós komplemens egy főnévi fej után áll.

A prepozíciós szintagmák érdekessége még, hogy a prepozíció gyakran elmarad (40). Ez különösen az újkeletű terminusokra vonatkoznak, amelyekről Béjoint és Ahronian (2008) azt állítja, hogy az angol nyelv hatására tűnt el belőlük a prepozíció, de az, hogy a főnevek sorrendje a francia szórendet követi, ez ellen szól.

(40) a. *code source*
kód forrás
'forráskód'

b. *accès* *Internet*
elérés internet
'internetelérés'

Ez azonban számunkra nem jelent problémát, hiszen az összetett főneveket felismerő minta ezeket is fel fogja ismerni, ugyanis ezen esetben egymás után álló főnevekről van szó.

4.2.2. Melléknevek helye a főnévi terminusokban

A melléknevek helye sokat elárul arról, hogy a terminus része-e vagy sem. A melléknév helye alapértelmezés szerint a főnévi fej után található, de a fej előtt is előfordulhat (ld. erről az 4.1.2. alfejezetet). Az egyik ilyen eset az, amikor az adott melléknevet

hangsúlyozzuk, ebből következően a tárgyilagosságot megkívánó szaknyelvben ez nehezen képzelhető el. Ezt szemlélteti a (41)-ben lévő szerkezet is:

(41) a. *un* *réseau* *filaire*
ART-INDEF.MASC.SG hálózat vezetékes
'vezetékes hálózat'
b. **un filaire réseau*

Ezenkívül, a négy melléknévcsoport (csoportosító, nem csoportosító, relációs melléknevek és sorszámnevek) közül nagyobbrészt azok lehetnek terminusok részei, amelyek a főnévi fej alfaját jelölik. A négy közül ez a tulajdonság a csoportosító és relációs melléknevekre igaz, amelyek csak a főnév után állhatnak: ezt is szemlélteti a (41)-ben szereplő példa, ahol a melléknév relációs. A gyakran használt, egyszótagú melléknevek szintén kevés eséllyel válhatnak a terminus részévé, de előfordulhatnak azok módosítójaként (42a). Ugyanez igaz a képzett intenzionális melléknevekre (pl. *soi-disant*), amelyek állhatnak terminusok előtt is, de nem képezik azok szerves részét (42b).

(42) a. *un* *grand* *réseau* *filaire*
ART-INDEF.MASC.SG nagy-MASC.SG hálózat vezetékes
'nagy vezetékes hálózat'
b. *un* *prétendu* *réseau* *filaire*
ART-INDEF.MASC.SG állít-PART.PAST.MASC.SG hálózat vezetékes
'egy állítólagos vezetékes hálózat'

Korábbi vizsgálatunkban (Nagy 2009c) azt tapasztaltuk, hogy egy informatikai jellegű szövegeket tartalmazó korpuszban nem szerepelt olyan terminus, amely melléknévvel kezdődött volna, azonban ez kisebb számban előfordulhat más típusú korpuszokban (43).

(43) a. *petite* *aiguille*
kis-FEM.SG mutató
'kismutató'
b. *premier* *ministre*
első-MASC.SG miniszter
'miniszterelnök'

Saját tanulmányunkon kívül nem rendelkezünk mért adatokkal arról, hogy a terminusok hány százaléka kezdődik valamilyen melléknévvel, ezért az előbb említett vizsgálat eredményei alapján nem vesszük figyelembe a saját terminológiakivonatoló esetében azt a lehetőséget, hogy egy melléknév egy terminus előmódosítója lehessen.

4.2.3. Determinánsok és egyéb adjunktumok a terminusokban

A legegyszerűbb esetben a főnévi terminus állhat egy darab főnévből, például *narcolepsie* 'narkolepszia'. Ezen kívül, ahogy a 4.2. eddigi alfejezeteiben bemutattuk, rendelkezhetnek

(legtöbbször determináns nélküli) prepozíciós komplementumokkal és (legtöbbször a főnévi fej mögött álló) melléknévi adjunktumokkal is. A 4.1. fejezetben felsorolt komplementumok és adjunktumok közül tagmondat nem fordulhat elő terminusokban, az csak a köznyelvi főnévi csoportokra értendő.

Tagmondatok előfordulhatnak ugyan főnévi terminusokban, de azok már nem képezik a terminusok részét, ugyanúgy, ahogy az előttük álló determináns(ok) sem. Ez több tényezőből is következik. Az egyik ilyen tényező a 3.2. fejezetben leírt, a terminusok létrehozásának egyik alapelvéből fakad: a sageri (1990) három fő kritérium a (1) gazdaságosság (*economy*), (2) pontosság (*precision*), (3) megfelelőség (*appropriateness*). A (1) gazdaságosság azt jelenti, hogy a terminusnak nem kell feltétlenül új szónak lennie: a köznyelvben használatos szavak vagy azok kombinációja is könnyen válhat terminussá, a (2) pontosság fogalma azt takarja, hogy egy terminus legyen elég pontos, hogy ne legyen kétértelmű, a (3) megfelelőség pedig a két korábbi kritérium konszolidációja: egy terminus legyen eléggé gazdaságos, de pontos is, ezáltal lesz megfelelő is. Ezen felül, egy terminust csak egyféleképpen lehet használni, szinonimája elvileg a wüsteri definíció alapján sem lehet.

Kevésbé valószínű, hogy a vonatkozói mellékmondatot tartalmazó terminusokat mindenki mindig ugyanúgy ismételné, ráadásul nem is felel meg a megfelelőségi kritériumnak. Erre példaként szolgálhat az alábbi szabadalmi idézet:

(44) *La poudre, qui est hygroscopique, ...*
ART-DEF.FEM.SG por aki/ami/amely van higroszkópos
'A por, ami higroszkópos, ... '

Ez a terminusjelölt így nem megfelelő, de mint 'higroszkópos por' felkerülhet a terminuslistára, a közbülső elemek nélkül. Ekkor viszont a teljes NP valóban nem felel meg a főnévi terminusnak.

A terminust bevezető determinánsok nem részei az előbbinek, mert a determináns figyelembevétele a terminus definíciójával több okból is összeférhetetlen lenne. Egyrészről a terminus csak fogalmat jelöl – azaz konkrét példányt nem – így a referenciális értelmű determinánsok sem lehetnek részei a terminusnak. Ráadásul a determinánsok jelenléte a megfelelőségi kritériumnak sem felel meg, mert a determinánsok nélkül is ugyanazt a fogalmat denotálják. Ezen kívül a terminológiai adatbázisokba – ugyanúgy ahogy a köznyelvi szótárakba is – a főnevek nem determinánssal kerülnek be. Ezért valójában nem a szintaktikai értelemben vett NP-knek, hanem egy közbülső kategóriának, az N'-nak felelnek meg a főnévi terminusok.

4.3. A többszavas francia terminusok belső szintaktikai jellemzői

Az összetett terminusok általában meghatározott belső szerkezettel rendelkeznek, amelyek az adott nyelvtől függnek. Az angolban ez az egyszavas főnév (pl. *thrombosis* 'trombózis'), vagy főnév és főnév kombinációja (pl. *dialog box* 'párbeszédablak'), vagy melléknév és főnév kombinációja (pl. *central heating* 'központi fűtés').[15] L'Homme (2004) szerint a francia nyelvben a leggyakoribb a főnév+melléknév és a főnév+prepozíció+főnév hármas. Egy korábbi tanulmányunkban (Nagy 2009c) egy informatikai korpuszból gyűjtöttünk ki terminusokat, amelyeket a belső szerkezetük alapján kategorizáltuk, majd azokból egy rangsort állítottunk össze az összetétel gyakorisága szerint, ezt a 4.1. táblázat szemlélteti:

4.1. táblázat: Terminusok szerkezetének gyakorisága[16]

Belső szerkezet	Arány	Példa
N	37,45%	*conteneur* 'konténer'
N N	20,21%	*retour chariot* 'kocsivissza'
N P N	14,47%	*fin de ligne* 'sorvége'
N A	11,91%	*nombre hexadécimal* 'hexadecimális szám'
N P D N	5,53%	*formatage du texte* 'szövegformázás'
N P N A	1,91%	*bouton de soumission spécialisé* 'testreszabott Küldés gomb'
N N N	1,49%	*paire nom valeur* 'név érték pár'
N P N P N	1,06%	*note de fin de page* 'lábjegyzet'
N P D N N	0,85%	*création des pages web* 'weboldalak létrehozása'
N P N N	0,64%	*création de pages web* 'weboldalak létrehozása'
N ADV A	0,64%	*lien déjà visité* 'már látogatott link'
N A ADV	0,64%	*texte aligné à droite* 'jobbra igazított szöveg'
N P N P D N	0,43%	*mise en ligne du site* 'webhely formázása'
N P D N A	0,43%	*codage des caractères spéciaux* 'speciális karakterek kódolása'

A kapott eredmények többé-kevésbé tükrözik a francia nyelvű terminusok belső szerkezetével kapcsolatos gyakorisági értékeket (pl. L'Homme 2004), ahol a leggyakoribb az egyszavas főnév, az egymás mellett álló főnevek, a főnév+prepozíció+főnév, valamint a főnév+melléknév kombinációk.

A mintával való illesztés hátránya, hogy – igaz csak kis arányban – de mindig előfordulnak olyan minták, amelyek nem szerepelnek a felsorolásban. A 4.1. táblázat is csak a leggyakoribb 14 mintát tartalmazza. Ilyen kivételre példa Nagy (2009a) szerint a

[15] A francia nyelvű terminusok belső szintaktikai összetételéről viszonylag kevés publikáció érhető el, így ebben a fejezetben L'Homme (2004)-re és egy saját publikációra támaszkodunk (Nagy 2009c). Hasonló témában a legtöbb publikáció az összetett szavak tipológiáját taglalja (a többszavas terminusok nagy része is összetett szó), ilyen például Gross (1996), Silberztein (1990) vagy Mathieu-Colas (1996).

[16] A fenti minták gyakoriságának vizsgálatakor minden terminusnak csak egyetlen előfordulását vizsgáltuk.

relation est-un ’ISA kapcsolat’ vagy ’általánosítás’ vagy a *relation un-à-un* ’egy-az-egyhez kapcsolat’, amelyeknek mintái rendre főnév-ige-névmás és főnév-névmás-prepozíció-névmás.

Az utóbbi tanulmányunkban (Nagy 2009c) azt is vizsgáltuk, hogy a terminusok belső szerkezetéhez milyen valószínűségi értékeket rendelhetünk. Ez azt jelenti, hogy egy adott belső szerkezethez megállapítottuk azt is, hogy az milyen valószínűséggel terminus, tehát minden egyes gyakori mintánál kiszámítottuk, hogy a mintához tartozó konkrét előfordulások az esetek hány százalékában voltak terminusok. Ezt szemlélteti a 4.2. táblázat, amelyben minden egyes mintához nemcsak a valószínűségi értékeket rendeltük, hanem adtunk is példát arra vonatkozóan, hogy egy adott mintához milyen terminus és nem terminus megnyilvánulások tartozhatnak.

4.2. táblázat: A minták terminusi valószínűsége

Minta	Terminus	Nem terminus	Valószínűség
N PREP N PREP N PREP N	*système de gestion de base de données* (adatbáziskezelő-rendszer)	-	100%
N PREP N A N	*langage de programmation orienté objet* (objektumorientált programozási nyelv)	-	100%
N A N	*programmation orientée objet* (objektumorientált programozás)	-	100%
N N	*ramasse-miettes* (szemétgyűjtő) *navigateur web* (webböngésző)	*mot clef* (kulcsszó)	98%
N N N	*programmation côté serveur* (szerver oldali programozás)	*mot clef interface* (interfész kulcsszó)	86%
N PREP N A	*nombre en virgule flottante* (lebegőpontos szám) *hiérarchie de classes unique* (egyedi osztályhierarchia)	*partie de code utile* (hasznos programrészlet)	75%
N PREP N	*durée de vie* (élettartam) *chaîne de décision* (döntési lánc)	*point de vue* (nézőpont)	75%
N	*disque* (lemez) *objet* (objektum)	*arbre* (fa) *cas* (eset)	70%
N A	*classe abstraite* (absztrakt osztály) *héritage multiple* (többszörös öröklődés)	*module requis* (megkívánt modul) *machine lente* (lassú gép)	38%
N PREP INF	*code à exécuter* (végrehajtandó kód)	*capacité à étendre* (kiterjeszthető kapacitás)	10%

A fenti táblázat alapján látható, hogy vannak olyan terminusszerkezetek, amelyek nagyobb

valószínűséggel terminusok belső szerkezetei. Ez nagyban megkönnyíti a terminusok automatikus kivonatolását, mert ezek alapján látható, hogy ha a szövegben egy adott mintának megfelelő karaktersorozat található, az is megállapítható, hogy az milyen valószínűséggel válik terminussá. Ha például egy főnév-prepozíció-főnév-melléknév-főnév kombinációra bukkanunk, akkor az szinte biztosan terminus. Az adott mintán belüli valószínűség nem egyezik meg a tényleges előfordulási valószínűséggel: a három egymás utáni főnév ugyan nagy valószínűséggel terminus (4.1. táblázat), de ritkán fordul elő: a 4.2. táblázatból látható, hogy az N-N-N minta csak 1,5%-os valószínűséggel szerepel. A leggyakoribb minták között szereplő egyszavas főnév vagy főnév és melléknév páros ennél rosszabb eredményt mutat.

5. A terminológiakivonatolás módszerei

Az 5. fejezet a disszertáció magját képezi, ismerteti a TE általános lépéseit. A 5.1. alfejezetben felsoroljuk a TE céljait, ami alapján meg tudjuk határozni, melyek azok az algoritmusok, amelyekre szükség van egy terminológiakivonatoló fejlesztésekor. A következő rész (5.2. alfejezet) célja azoknak a lépéseknek az ismertetése, amelyek a TE során nélkülözhetetlenek: e részben foglaljuk össze, milyen feladatokat milyen sorrendben kell végrehajtani annak érdekében, hogy ezen tudást a későbbiekben fel lehessen használni a saját terminológiakivonatoló létrehozásakor. A terminusjelölt-lista létrehozása és szűrése valójában a 5.2. alfejezetben közölt lépések közül a legfontosabb és a legkritikusabb: ezek azok a feladatok, amelyek a leginkább hatással vannak a terminológiakivonatolók hatékonyságára. A terminusjelölt-lista felállítására és szűrésére számos metódust dolgoztak ki. Mivel mindkét szakasz kritikus, ezért ezeket külön tárgyaljuk a következő két alfejezetben (5.3. és 5.4.). A metódusokat nem az alapján csoportosítjuk, hogy melyik mire alkalmazható, hanem aszerint, hogy melyek szabály alapúak és melyek statisztikai módszerek. A 5.3. alfejezet a szabály alapú modulokat tartalmazza, a 5.4. alfejezet a statisztikai módszereket írja le.

Ahogy a szabály alapú és a statisztikai módszereket is részletező alfejezetekből kitűnik, igen sok metódust dolgoztak ki terminusok kinyerésére, ezért nem elegendő azokat csupán felsorolni – tudnunk kell, hogy ezek közül a saját kivonatoló létrehozása során melyeket tudjuk felhasználni. Erre szolgál az 5.5. alfejezet, amelyben összefoglaljuk a különböző terminológiakivonatoló módszerek sajátosságait. Ezt követi az 5.6. fejezet, amelyben a francia nyelvre készült terminológiakivonatolókat foglaljuk össze. Az utolsó, 5.7. alfejezetben az általunk kidolgozott terminológiakivonatoló főbb paramétereit mutatjuk be.

5.1. A terminológiakivonatolás célja

A terminológiakivonatoló alkalmazások eszközeinek és moduljainak bemutatása előtt először a TE céljait határozzuk meg. Erre azért van szükség, mert mint ahogy azt a későbbiekben is látni fogjuk, annak a tényezőnek, hogy milyen célra készül a terminológiakivonatoló eszköz, már nagy szerepe van abban, hogy mely algoritmusokat választjuk. Cabré és mtsai (2001) szerint nem ugyanazokat a technikákat kell használnunk,

ha a terminusok kinyerését terminológiai adatbázis létrehozására vagy dokumentumok automatikus indexelésére használjuk.

Az alábbiakban azokra a főbb területekre térünk ki, amelyekhez már hoztak létre terminológiakivonatoló alkalmazást. Azt is leírjuk, hogy ezekben mennyire szükséges a terminusok pontos kinyerése, ám számos más célja is van a terminológiakivonatolásnak. A felsorolt célokon túl léteznek más típusú alkalmazások, például olyanok, amelyek a terminológiakinyerést a gépi fordítás céljába próbálják állítani (pl. Vasconcellos 2001), illetve olyanok is, amelyek ezt információkinyerő célokra használják (pl. Ahmad 2001).

5.1.1. Fordítói munka elősegítése

Ha valaki már dolgozott olyan fordítási munkákban, amelyekhez egy egész csapatra szükség van, különösen átérzi annak szükségét, hogy a fordítás megkezdése előtt minden résztvevőnek legyen a kezében egy olyan lista, amely tartalmazza az összes terminus célnyelvi megfelelőit. Fordítói csoportban akkor szokás dolgozni, ha nagy terjedelmű szakmai szöveget (mondjuk egy szoftver vagy elektronikai termék leírását) nagyon rövid idő alatt kell lefordítani. De természetesen a fordításon nem szabad látszódnia, hogy több különböző ember dolgozott rajta, aki ugyanazt a *terminus technicus*t esetleg másképp fordítja, mint a többiek. Ez viszont folyamatos egyeztetést igényel a csoport tagjai között, ami jelentősen csökkenti a munka hatékonyságát. Ekkor nagy segítséget jelentene, ha az adott szövegből először kinyernénk automatikusan a terminusokat, majd azoknak előre megszerezzük a célnyelvi megfelelőit. Ez a lista akkor is fontos lenne, ha nem fordítói csoportban dolgozunk, hanem egyénileg.

A Kis B. (2005) által kidolgozott alkalmazás kifejezetten erre épít, sőt, már magát a terminus fogalmát is minden olyan elemre kiterjeszti, amelyet mindig ugyanúgy kell fordítani. Ez persze sok szókapcsolat esetén is igaz lehet, ami nem is terminus, de erre a célra olyan alkalmazást kell készíteni, amely a terminusokat nagy fedéssel nyeri ki. Ez azt jelenti, hogy nem baj, ha a terminuslistában sok olyan elem van, amire nincs szükség, de lehetőleg az összes olyan szó vagy szókapcsolat benne legyen, amit mindenképpen azonos módon kell fordítani, hogy később emiatt már ne kelljen a fordításban résztvevőknek egyeztetniük. Azonban a fordítói munka elősegítésénél nem feltétlenül az a fontos, hogy terminusok szerepeljenek a listában, hanem azok a szavak/kifejezések, amelyeket egységesen kell fordítani.

5.1.2. Dokumentumok indexelése

A dokumentumindexelés azt jelenti, hogy kiválasztjuk az adott dokumentumból azokat a szavakat és kifejezéseket, amelyek az adott szöveget minél jobban leírják. Erre akkor van szükség, ha később ezek közül a dokumentumok között szelektálnunk kell: például, ha csak bizonyos biológiai szövegekre van szükségünk, akkor a dokumentumhalmaz címkéi alapján kiválasztjuk a megfelelő szövegeket. Ilyenek az internetes keresőmotorok is (például a Google vagy az AltaVista), amelyek a megadott keresési paraméterek alapján kiválasztják azokat a dokumentumokat, amelyekre szükségünk van. Ehhez először a keresőmotoroknak minden egyes új dokumentumot indexelniük kell a bennük található, főleg főnévi szavak vagy szócsoportok alapján.

Az, hogy a TE mennyire szorosan összefonódik a dokumentumindexeléssel, mi sem bizonyítja jobban, hogy a Ruslan Mitkov által szerkesztett számítógépes nyelvészeti „biblia" terminológiakivonatolással foglalkozó fejezete (Jacquemin és Bourrigault 2003) együtt kezeli a két típusú alkalmazást. Cabré és mtsai (2001) a terminológiakivonatolók eredményeinek összehasonlításakor olyan alkalmazásokat is vizsgál, amelyeket kizárólag dokumentumindexelési célokra használnak. Léteznek olyan terminológiakivonatoló alkalmazások is, amelyek nem dokumentumindexelési célokra készültek, mégis olyan algoritmusokat használnak, amelyek főleg csak erre alkalmasak. Ilyen Lefever és mtsai (2009) terminológiakivonatolója is, amely az indexelésben gyakran használt *weirdness*-algoritmusra épül.

A dokumentumindexelés számára a terminusok azok az elemek, amelyek egy szöveg tartalmát a legjobban leírják. Így itt az a fontos, hogy minél több kifejezést gyűjtsünk, amelyek az adott szöveget könnyen elérhetővé teszik a keresés szempontjából. Ezen alkalmazásoknál a pontosság nem számít annyira, mert vannak olyanok, amelyek például csak egyszavas terminusjelölteket keresnek (pl. Ahmad és mtsai 1999). A lényeg, hogy olyan elemek legyenek a listában, amelyekre a felhasználói kereső lekérdezés a lehető legrelevánsabb dokumentumokat adja vissza.

5.1.3. Terminológiai adatbázisok létrehozása

Feltehetően ez az a cél, amely a legnagyobb pontosságot igényli a terminológiakivonatoló alkalmazásoktól: csak az adott szakterület terminusai kerüljenek be egy ilyen adatbázisba, és ha valamilyen írott korpusz a gyűjtés alapja, akkor az adatbázisban legyen benne az összes, korpuszban is előforduló szakkifejezés. Az ilyen adatbázisok lehetnek

többnyelvűek, ezáltal szótárként is használhatók, erre példa az autóipari terminusok francia, holland és angol nyelvű változatai (Lefever és mtsai 2009). Lehetnek egynyelvűek is, amelyek kiindulási pontjai lehetnek egy egynyelvű szakszótár létrehozásának is: a kivonatoló alkalmazás által megtalált terminusoknak a definícióit és jelentéseit kell csak kikutatni. Léteznek azonban olyan kivonatoló alkalmazások is, amelyek nem csak a terminusokat, hanem azok definícióit is képesek megtalálni egy adott korpuszban (pl. Piao és mtsai 2008).

Ezenkívül olyan alkalmazások is léteznek, amelyek nem csak arra vállalkoznak, hogy a terminusokat kinyerjék, hanem arra is, hogy azokat ontológiákba szervezzék (pl. Yirong és mtsai 2008). Ez azt jelenti, hogy a terminusok itt már hierarchikus viszonyba szervezettek attól függően, hogy az általuk denotált fogalom a többivel milyen viszonyban áll, amely nem csak alá-, mellé- és fölérendeltségi viszonyt ír le, hanem egyéb kapcsolatok feltárását is tartalmazza.

5.2. Terminológiakivonatolás lépései

Először bemutatjuk, hogyan működnek általában a terminológiakivonatoló alkalmazások. Ezt a részt azért emeljük ki a konkrét jellemzések előtt, mert vannak olyan lépések, amelyek minden kivonatolónál azonosak. Sőt, a terminológiakivonatás ezen műveleti lépései nemcsak azonosak, hanem kötelező érvényűek is, így a későbbiekben így oldjuk meg a terminológiakivonatolást. Cabré és mtsai (2001) és Sauron (2002) szerint a TE főbb lépései a következők:

A, terminusjelöltek kinyerése, amely egy adott korpuszból történik

B, ezen terminusjelöltek szűrése

C, terminusjelöltek validálása

D, a validált terminusok ontológiákba vagy glosszáriumba, adatbázisba történő felvétele

Mi első lépésként felvettük a korpusz, illetve nyelv kiválasztását is, hiszen a későbbiekben ennek is fontos szerepe lehet: nem mindegy, hogy a korpusz milyen szakterülethez tartozik, és az sem, hogy milyen nyelvű, mert lehet, hogy bizonyos nyelvekhez vagy másfajta nyelvi megközelítés szükséges, vagy más nyelvfeldolgozó modulok állnak rendelkezésre.

A következő lépés a korpusz nyelvi feldolgozása. Az igaz, hogy a nyelvi feldolgozás 100%-osan nem megbízható, de szinte minden terminológiakivonatoló használ

nyelvészeti információkat is – például azért, hogy csak a főnévi terminusokat nyerjék ki. A nyelvi információk használata Cabré és mtsai (2001) szerint történhet a terminusjelölt-lista felállításakor, illetve szűrésekor (vagy mindkettő folyamán is), de az biztos, hogy valamikor ilyen információkra is szükség van.

A terminusjelöltek kinyerésére és szűrésére két fő módszer létezik, a szabály alapú és a statisztikai módszer. A leggyakrabban használt mégsem ez a kettő, hanem ezek kombinációja, amelyet hibrid módszernek is nevezünk. Ez azt jelenti, hogy a hibrid alkalmazásokban mind szabály alapú, mind statisztikai módszereket is alkalmaznak. Cabré és mtsai (2001), valamint Ha és mtsai (2008) szerint a hibrid alkalmazások először a terminusjelöltek kinyerésére a statisztikát alkalmazzák, majd azok szűrésére nyelvi filtereket. Ugyanakkor mind a statisztikai, mind a nyelvi modulok akár az első terminuslista létrehozásakor, akár a szűrés folyamán is megjelenhetnek, így a későbbiekben ezen módszereket inkább nem az alapján csoportosítjuk, hogy mely folyamatban vesznek részt, hanem az alapján, hogy melyek a szabály alapú és melyek a statisztikai elemek. Mivel ezekből elég sok van (nagyon sok terminológiakivonatoló alkalmazás készült már), ezért ezeket külön alfejezetben ismertetjük.

Ezt követi a validálás folyamata, amelynek során az dől el, hogy az adott program milyen hatékonysággal működik. Ez főleg kétféleképpen dönthető el: (1) szakértőket kérnek arra, hogy a kivonatolt terminusok egy részéről megállapítsák, azok valóban terminusok-e; (2) a program összehasonlítja kimenetét egy terminológiai adatbázissal.

A Sauron (2002) által is említett glosszáriumok és ontológiák felállítása munkánk során nem cél, így azzal nem foglalkozunk bővebben. Ez akkor lehet fontos, ha nemcsak a terminusokat szeretnénk megvizsgálni, hanem azt is, hogy azok milyen viszonyban állnak egymással: mely terminus mely másik terminusnak a szinonimája, hiperonimája vagy hiponimája.

5.2.1. Korpusz/tudományterület kiválasztása

Az első lépés minden esetben a korpusz kiválasztása, amellyel együtt a tudományos területet is meghatározzák. A korpusz kiválasztása azért elsődleges, mert sok esetben eldől már itt a kivonatolási módszer is. Sok terminológiakivonatolót úgy terveznek, hogy csak egy bizonyos területre koncentráljanak. Például Hoste és mtsai (2008) korpuszként kizárólag betegek kórlapjait/orvosi jelentéseit vizsgálják, és abból nyernek ki bizonyos orvosi/biológiai terminusokat. Brekke és mtsai (2006) a közgazdaságtan és az

államigazgatás területére korlátozza kutatásait, azon belül egy inkább tudományos jellegű adatbankot használ fel. Anstein és mtsai (2006) a kémiával, vegytannal kapcsolatos kifejezések kigyűjtését végezte el.

Az egy területre korlátozódó terminológiakivonatolóknak sok előnyük és hátrányuk van. Ezen alkalmazások azért lehetnek jobbak, mint az általános terminuskivonatolók, mert megvalósításuk a terület specifikussága miatt egyszerűbb és ezért hatékonyabb is. Észrevehető, hogy ugyan a terminusok belső szerkezeteiben vannak hasonlóságok a különböző domének között, de minden esetben vannak speciális jegyek is, amelyek megkönnyítik azt, hogy az adott területen a terminológiakivonatoló még nagyobb hatékonysággal működjék. A kémiában vagy a vegyészetben a különböző ionok, molekulák, gyökök elnevezései egy sémát követnek, ezáltal így is könnyen megkülönböztethetők. Az elnevezések ezen a tudományterületen rendelkeznek egy belső összetétellel is, azaz az affixumok vizsgálata előnyt jelent, ugyanis vannak bizonyos prefixumok (pl. *iso-, hydro-* stb.), illetve szuffixumok (pl. *-id, -on* stb.), amelyek csak ezen területre jellemzőek, így már a különböző affixumok keresése is jó eredményt hozhat. Azonban könnyen belátható, hogy más tudományágakban (például műszaki területek vagy az államigazgatás) ezek vizsgálata nem hoz számottevő eredményt.

A szakterületre korlátozott terminológiakivonatolásnak más előnyei is vannak: ugyan más területre nem alkalmazhatók, de az adott tudományágban nagyon pontosan működnek. Ezeket általában nem arra használják, hogy bármilyen korpuszban jól működjenek, hanem arra, hogy minél pontosabban feltérképezzék az adott terület terminológiáját és abból például adatbázist készítsenek.

Ezzel szemben egy általános terminológiakivonatoló bármely szakterületen működik, de ebből adódóan kisebb pontossággal, mivel nem tudja kihasználni egy adott tudományág terminológiájának specifikus jellemzőit. Ezt a jelenséget például jól jellemezheti az *or gate* (’VAGY-kapu’) kifejezés (Savary 2000), amelyet egy számítógép-architektúrákra betanított terminológiakivonatoló könnyedén felismerhet, de ha az *or* az adott szövegben szóközzel van elválasztva a *gate* szótól, akkor egy általános ilyen típusú alkalmazás az *or* szót könnyedén tekintheti csak egy kötőszónak, így ez a kifejezés lehet, hogy nem kerül be a terminuslistába.

A korpusz kiválasztásakor a nyelv kiválasztása is megtörténik, ugyanis nagyon ritka az olyan alkalmazás, amely bármely nyelvre jól működne. A nyelv kiválasztása más szempontból is érdekes; a legtöbb terminológiakivonatoló ugyanis nyelvi modulokat is

alkalmaz, például szófaji címkézőt vagy szótövesítőt. Ezen modulok esetében kérdéses, hogy az adott nyelvre milyen szinten készült már el ilyen alkalmazás, hiszen ez nagyban befolyásolja azt, hogy ezeket milyen megbízhatósággal alkalmazhatjuk a TE során. A nagyobb világnyelvekre már számtalan, akár nyílt forráskódú, akár licensz megvásárlásához kötött modul érhető el, a kisebb vagy ritkább nyelvekre viszont kevesebb, amelyek hatékonysága lehet rosszabb is.

5.2.2. A korpusz szegmentálása, nyelvi elemzése

Ha már megvan a vizsgálandó korpusz, amelyből ki szeretnénk nyerni a terminusokat, akkor azt először szegmentálni kell. A mondatokra és tokenekre történő szegmentálást legjobban Mikheev (2003), a szófaji címkézést Voutilainen (2003) írja le, a szintaktikai elemzést Carroll (2003), így ezekben a fejezetekben ezekre építünk. Fontos megjegyezni, hogy az első három szakaszt sok program egyszerre el tudja végezni, így nem szükséges ezen modulokat külön beilleszteni, hanem elég csak egyet, ami mindhármat képes elvégezni.

5.2.2.1. Korpusz szegmentálása mondatokra

Először a bemeneti korpuszt mondatokra kell bontani, mert ez elengedhetetlen a szófaji címkézéshez, szótövesítéshez, és a mondat szintaktikai elemzéséhez. A mondatokra történő szegmentálásnak több módja van, de általában itt is két alaptípus különböztethető meg, mint a legtöbb számítógépes nyelvészeti alkalmazások esetében: ezek a szabály alapú, illetve a statisztikai módszerek. A szabály alapú módszerek rendkívül egyszerű alapelven működnek: az egyik legismertebb és leginkább egyszerű megoldás a mondathatárok bejelölésének megvalósítására az úgynevezett „mondatvégi írásjel-szóköz-nagybetű” algoritmus, amely a fenti hármasok közepénél bejelöli a mondathatárt. Ez azonban nem mindig elegendő, mert idézőjelek, valamint zárójelek is közbeékelődhetnek ebbe a mintába, így a fenti reguláris kifejezést át kell írni, hogy ezeket is kezelni tudja: így kapjuk a [.?!][ )”]\s[A-Z] reguláris kifejezést, ahol az \s akármilyen üres karakter lehet.

E módszer – mint sok más számítógépes nyelvészeti alkalmazás – nagyon egyszerűnek tűnik, de korántsem az. Az esetek nagyobb részében ez jól működik, de rengeteg esetben van a mondat közepén pont (például rövidítések után), amelyet nagybetű követ, ezt azonban nem tekinthetjük mondathatárnak. Ilyenkor egyéb szűrők alkalmazása is szükséges, például egy rövidítéslista.

A statisztikai módszerek esetében általában valamilyen előre bejelölt korpusz alapján dolgozik az alkalmazás, tehát ez utóbbi alapján statisztikai módszerekkel dönti el, hogy egy adott szövegben hol vannak a mondathatárok. Ezek működése általában nagyobb pontosságú, de annyira nem jelentős a különbség a szabály alapú módszerekhez képest, mert a 90%-os pontosságot és fedést mindkét módszerrel el lehet érni (Mikheev 2003).

5.2.2.2. Korpusz szegmentálása tokenekre

Mikheev (2003) szerint a token egyedi előfordulású szóalak, azaz egy olyan karaktersorozat, amelyet legtöbbször szóközök vagy írásjelek határolnak, és amely egy szövegben többször is előfordulhat. Azért nem szavakra történő bontás ez, mert a tokenek jelentős része ugyan szó, de nem csak az lehet, a *kutyát* mindkettő, de a *15000* csak token. Sőt, az automatikus tokenizáló programok esetében olyan írásjelek is tokennek minősülnek, amelyek nem képezik más szavak részét, például a pont, a vessző. Az általunk alkalmazott Machinese is tokennek veszi ezeket az írásjeleket.

A tokenizálás folyamatát is jelentősen megnehezítik a kivételek. Alapjában véve egy tokent szóközök, vagy egyéb írásjelek (mint például vessző, kettőspont, kötőjel, stb.) határolnak, azonban nem minden szóköz vagy kötőjel minősül tokenhatárolónak: erre példa a könnyebb olvashatóság kedvéért írt *16 000* vagy 3*0 %*, vagy ha az angol nyelvet vesszük alapul, akkor a *twenty-six* egy token, de a *San Diego-based* kettő, ráadásul nem a szóköznél, hanem a kötőjelnél kell azt szétválasztani.

5.2.2.3. Tokenek szófaji címkézése és lemmatizálása

A tokenizálás folyamata után történhet meg a tokenek szófaji címkézése és lemmatizálása. A szófaji címkézés (*POS-tagging*) abból áll, hogy a tokeneket megcímkézzük azok szófajával, esetleg egyéb inflekciós tulajdonságaival, azaz megtudjuk, hogy az adott token milyen szófajú szó az adott kontextusban, és milyen inflekciós tulajdonságokkal rendelkezik. A magyar nyelv esetén a *kutyát* token címkéjének az alábbi adatokat kell tartalmaznia: ez egy főnév, amely egyes számban áll és tárgyraggal rendelkezik.

Egy adott természetes nyelvben egy adott szónak több szófaja is lehet, így a szövegkörnyezet figyelembe vétele nélkül ritkán lehet jó egy szófaji címkéző. Erre jó példa az angol nyelv, ahol sok token lehet egyszerre főnév vagy ige is: az adott szerepet mindig az adott kontextus határozza meg, amelyben ez az elem található. Ilyen szóalak a *target*, amely főnévként ’cél’, igeként ’megcéloz’ jelentésű, vagy a *dog* szó, amely főnévként

’kutya’, igeként ’követ valakit’ jelentésű. Hasonló jelenséggel találkozhatunk a francia nyelvben is: Plante és Dumas (1998) szerint a francia szavak 30%-a több szófaji kategóriához tartozhat. Például a *table* szó egyszerre lehet főnév ’asztal’, valamint egyes szám jelen idejű ige ’alapoz valamire’ jelentésben, illetve a *manger*, amely főnévként ’étel’, főnévi igenévként ’enni’ jelentésű. Ilyen esetekben a szófaji címkéző programnak el kell döntenie azt is, hogy a kettő közül melyik szófajhoz tartozik az adott token, így a szófaji címkézést gyakran nevezik szófaji egyértelműsítésnek is.

A lemmatizálás szótövesítést jelent, azaz adott egy token, annak szófaja, és ezen adatokból a program megállapítja, hogy mi az adott token ragok nélküli szótöve. A lemmatizálás nem azonos a *stemming* folyamatával: a lemmatizálás csak az inflekciós affixumokat távolítja el, egy *stemmer* a derivációs affixumokat is. A *barátkozásait* token lemmatizált szótöve tehát a *barátkozás*, míg *stemmelt* töve a *barát*.

Azt, hogy ezek közül melyiket választjuk, előre el kell dönteni. A szófaji címkézés akkor előnyös, ha például dokumentumokat indexelünk tartalmuk alapján: ekkor mindegy, hogy a *demolition* főnév vagy a *demolish* ige fordul elő a szövegben; elég csak azt számon tartanunk, hogy a *demoli-* tő hányszor fordul elő. Terminológiakivonatolásnál viszont minden alkalmazás lemmatizálót használ, hiszen terminusok kinyerésekor fontos lehet a szóalak.

A szófaji címkézésre azért van szükség a TE során, mert előfordulhat, hogy minták alapján keresünk terminusgyanús főnévi csoportokat, például melléknév-főnév párosokat (például *operációs rendszer*, *off-line böngészés* stb.). A lemmatizálás azért fontos, mert ha azt is nézzük, hogy az adott terminus hányszor fordul elő a szövegben, akkor ne számítsuk külön előfordulásnak az *operációs rendszer*, *operációs rendszert*, *operációs rendszertől* stb. alakokat.

5.2.2.4. Szintaktikai elemzés

Léteznek olyan alkalmazások is, amelyek a már mondatokra bontott és szófajilag címkézett korpuszban a szintaktikailag egybetartozó részeket is bejelölik. Ez azt jelenti, hogy például összecsoportosítják azokat az elemeket, amelyek együttesen kitesznek egy főnévi, melléknévi, határozói stb. szintagmát. Ez a TE során azért fontos, mert ha főnévi terminusokat keresünk, akkor elég csak a bejelölt főnévi csoportokon végigmenni, és azokról eldönteni, hogy terminusok-e. Ezt azonban érdemes megfontolni, mert már maguk a szófaji egyértelműsítők is körülbelül 5%-os hibaaránnyal dolgoznak (L’Homme 2004),

így ha még a szintaktikai elemző hibaarányait is figyelembe vesszük, akkor már elég nagy lehet az eltérés. Ráadásul, mint ahogy már korábban említettük (4.2.3. fejezet), a terminusok nem egészen olyan NP-k, mint a köznyelvi NP-k: például nem tartalmazhatnak alárendelő vonatkozói mellékmondatokat (azaz nem lehet olyan szerkezetű, mint az a köznyelvi NP, hogy *az a szomszéd, amelyik állandóan panaszkodik*).

A szintaktikai elemzésnek két fő fajtája van: *deep parsing* és *shallow parsing*, vagy az utóbbi más néven *chunking* (mély, illetve lapos szintaktikai elemzés). Az első sokkal pontosabb: jelöli azt is, hogy például egy főnévi csoport melyik igének vagy prepozíciónak argumentuma, illetve azt is, hogy milyen szerepet tölt be az adott mondatban. Ezzel ellentétben egy *chunker* csak az alapvetőbb szintagmákat nyeri ki a szövegből, nem foglalkozik esetleg azzal, hogy abban milyen egyéb beágyazott, más vagy ugyanolyan típusú szintagmák is szerepelhetnek (Carroll 2003).

5.2.3. Terminusjelölt-lista összeállítása

A szófaji jelöléseket tartalmazó korpuszból már könnyebb kinyerni a terminusokat, mint egy egyszerű szövegből. Ebben a lépésben különböző algoritmusokkal ismernek fel a kivonatolók terminusjelölteket, tehát ennek a folyamatnak a végén egy olyan listát kapunk, amely, remélhetőleg, majdnem az összes terminust tartalmazza. Azonban a probléma az, hogy ez az első lista valószínűleg sokkal több elemet tartalmaz majd, ezért van szükség ezen lista későbbi szűrésére. Mint ahogyan a következő lépésben, itt is több módszerrel találkozhatunk: vannak szabály alapú, statisztikai, illetve az ezek kombinációiból felépülő hibrid módszerek is. Ezeket az 5.3. és 5.4. alfejezetben fejtjük ki bővebben.

5.2.4. Terminusjelölt-lista szűrése

Az első terminuslista nagy valószínűséggel sokkal több elemet tartalmaz, mint amennyi a valódi terminusok száma. Ennek az az oka, hogy akármilyen módszerrel is nyerjük ki az első terminusjelölteket, minden esetben találhatók olyan elemek, amelyek ugyanolyan szabályszerűséggel rendelkeznek, mint a terminusok.

Ha például azt vesszük alapul, hogy a terminusok gyakran szerepelnek egy adott korpuszban, és csak gyakoriságot mérünk, akkor olyan elemek is bekerülhetnek a listába, amelyek ugyan gyakran szerepelnek az adott korpuszban, de mégsem terminusok. Ilyenek lehetnek például a különböző határozószók (*gyakran, jól*), vagy gyakori főnevek (*gép*, *elem*). Ha azt vesszük alapul, hogy a terminusok bizonyos belső szerkezettel rendelkeznek,

például melléknév és főnév, akkor olyan szerkezeteket is kivonatol az alkalmazás, mint *első lépés*, *nagy kérdés* stb. Ezek szűrésére különböző modulok állnak rendelkezésre, amelyet a 5.3. és 5.4. pontban ismertetünk részletesen.

5.2.5. Validálás

A validálás az a szakasz, amely során meggyőződünk arról, hogy a kivonatolt terminusok tényleg azok-e vagy sem. Erre is több módszer létezik, de alapvetően kétféle lehet: teljesen automatikus vagy kézi ellenőrzés.

Ha a validálási folyamat automatikus, akkor az általában azt jelenti, hogy a kinyert terminusokat az alkalmazás összeveti egy előre készített listával, vagy ha az alkalmazás rendelkezésére áll a korpusz egy olyan annotált változata, ahol a terminusok is jelölve vannak, ezek megvizsgálása igen egyszerűvé válik. Igen gyakori az az eset is, amikor az elemzés végén kikerült terminusokat összevetik egy terminológiai adatbázis elemeivel (Deane 2005; Wermter és Hahn 2005). Ha a terminusjelölt megtalálható az adatbázisban, akkor azt elfogadják terminusnak, ha pedig nem, akkor az valószínűleg nem szakkifejezés. Ennek azonban az a hátránya, hogy ha az adatbázis nem elég friss, akkor előfordulhat, hogy nem tartalmazza az összes olyan terminust, amelyeket az alkalmazás kinyert, így a validálásnál az eredmény nem lesz megfelelő.

A másik módszer esetében „bírák" alkalmazásával történik ez a folyamat, amely azt jelenti, hogy egy vagy több embert megkérnek arra, hogy a kijelölt terminusok egy – lehetőleg reprezentatív – részét nézzék át, és döntsék el, hogy azok közül melyek tényleg terminusok, és melyek nem. Ha több bírát alkalmazunk, az is észrevehető, hogy az eredmények nem mindig egyeznek meg: ennek legtöbbször nem az az oka, hogy a bírák nem értenek az adott szakterülethez, csak például nem mindig tudják megmondani, hogy a szóösszetétel így egyben is terminus, vagy elég csak egy részét annak tekinteni. Kifejezetten ilyen validálási technikát ismertet Zhang és mtsai (2008) vagy Frantzi és Ananiadou (1999).

A végleges hatékonyság kiszámítására alkalmazott mértékek azonban minden alkalmazásban közös: ez a két érték a pontosság (*precision*) és a fedés (*recall*), amelyeket nagyon gyakran használnak számítógépes nyelvészeti alkalmazások hatékonyságának kifejezésére. Mindkét érték 0 és 1 közötti valós szám, ahol a 0 a legkisebb pontosságot, illetve fedést jelenti, az 1 pedig ezen értékekből a legjobbat. A két érték egymástól teljesen

független, lehet olyan alkalmazás, amely nagyon nagy pontosságú, de kis fedésű, illetve fordítva.

A pontosság a TE esetén azt mutatja meg, hogy a kivonatolt terminusjelölt-listában milyen arányban fordulnak elő olyan elemek, amelyek nem terminusuk, tehát a zaj szintjét mondja meg. Minél több felesleges elem van a listában, a zaj annál nagyobb, az alkalmazás pontossága pedig annál kisebb, azaz annál inkább közelít a 0 érték felé. A pontosság kiszámítására az alábbi képletet alkalmazzuk:

$$pontosság = \frac{1}{kivonatolt\ helyes\ terminusok\ száma} \cdot ebből\ valós\ terminusok\ száma$$

A pontosság azt a mértéket adja meg, hogy a kivonatolt terminusok közül milyen arányban vannak a valódi terminusok. Ha ez a két szám megegyezik, tehát az összes kivonatolt elem terminus, akkor a formulából is látszik, hogy 1-et fogunk kapni. Ha a valós terminusok száma nagyon elenyésző a kivonatolt terminusok számához képest, akkor az érték majdnem 0 lesz. Ha nincs kivonatolt terminus, akkor a pontosság nem értelmezhető.

Ezzel szemben a fedés azt mondja meg, hogy a valós terminusok közül mennyit talált meg az adott terminológiakivonatoló alkalmazás, tehát a csend mértékét adja meg. Ebben az esetben azt nem figyeljük, hogy mennyi felesleges elem van a listában, csak azt, hogy mennyi nincs benne azok közül, amelyeknek benne kell lenniük. Minél több elem hiányzik a terminusjelölt-listából, a fedés annál kisebb, azaz annál inkább közelít a nullához. A fedés kiszámítására az alábbi formulát használjuk:

$$fedés = \frac{kivonatolt\ helyes\ terminusok\ száma}{valós\ terminusok\ száma}$$

A fedés azt mutatja meg, hogy milyen arányban szerepel a listában az összes igazi terminus. Ha ez a két szám megegyezik, tehát a terminusjelölt-lista az összes valós terminust tartalmazza, akkor a formulából is látszik, hogy 1-et fogunk kapni. Ha pedig a helyesen kivonatolt terminusok száma nagyon elenyésző a valós terminusok számához képest, akkor az érték majdnem 0 lesz.

Mint ahogy a fenti definíciókból is látszik, egy terminológiakivonatoló alkalmazásnál mindkét tényező nagyon fontos, és nem elegendő csak az egyiket figyelembe venni. Például van egy szövegünk, amelyben 100 terminus van. Az alkalmazás 10 jelölttel tér vissza, amelyek közül mind terminus. Ekkor a fedés egytized, a pontosság viszont kereken egy. Ha ugyanebből a szövegből egy másik alkalmazás 200 jelölttel tér vissza, amelyből 50 a valóban terminusok száma, akkor a pontosság kevesebb, mint az előző esetben, mert most egynegyed, a fedés viszont több mint az előző esetben, mert az

most egyketted. A szakirodalomban is ezen értékeket tüntetik fel, ha az alkalmazásuk hatékonyságát kell megadniuk, így a későbbiekben mi is ezeket az értékeket fogjuk használni a validálási folyamat során.[17]

Van lehetőség a két mérték összevonására is: ez az F-érték, amelyet a számítógépes nyelvészetben és a matematikai statisztikában gyakran használnak a hatékonyság mérésére. Ez az érték a pontosság és a fedés értékének harmonikus közepe (de nem átlaga). Így a hagyományos F-érték kiszámításának módját az alábbi formula adja meg:

$$\text{F-érték} = \frac{2 \cdot \text{pontosság} \cdot \text{fedés}}{\text{pontosság} + \text{fedés}}$$

Az előző példákban az első esetben 0,18 ez az érték, míg a másodikban 0,33. Ez a két érték nagyon jól megmutatja, hogy a második eset valóban jobb, mert a pontosság és a fedés kiegyensúlyozottabb, de az F-érték viszont elsimítja azt, hogy az alkalmazás melyik értéke az, amelyen változtatni lehetne.

5.3. A terminológiakivonatolók szabály alapú moduljai

Ebben a fejezetben azon módszereket soroljuk fel, amelyek a szabály alapú megközelítést jelképezik. A nyelvi alapú módszerek a terminológiakivonatolásnak akár a terminusjelölt-lista felállításában, akár annak szűrésében is szerepelhetnek.

5.3.1. Mintaillesztés

A mintaillesztés tekinthető a TE egyik legalapvetőbb módszerének: már az első hivatalos terminológiakivonatolók (pl. Plante és Dumas 1989) is ezt a módszert alkalmazta. Ennek lényege, hogy észrevehető egy belső szerkezet, amellyel a terminusok általában rendelkeznek, ezt a belső szerkezet nevezzük mintának. Ezek felsorolásával, vagy ezek ismeretében már a terminusok nagyobb része megtalálható.

Egy egyszerű példával szemléltetve, elég csak megadnunk olyan közismert mintákat, mint főnév-főnév (például az angol *dialog box* vagy *IP address*), vagy melléknév-főnév (például *operational system*). A mintaillesztés azonban erősen nyelvfüggő, mert ugyanaz a minta nem lehet minden nyelvre releváns. Ezt jól szemlélteti

[17] Az automatikus validálás során (amikor a kimenetet egy terminológiai adatbázissal vetjük össze) nem lehet a fedés értékét kiszámolni, csak a pontosságét, mivel annotált korpusz nélkül nem tudhatjuk, hogy valójában mik a szöveg valódi terminusai. Az automatikus validálással nem kapunk tényleges képet egy terminológiakivonatoló alkalmazás valódi eredményeiről. Ez olyan esetben hasznos, amikor például egy szövegből kinyert biztos terminusokon egyéb vizsgálatokat végzünk el, például megnézzük azok szövegkohéziós értékeit (pl. Deane 2005). Ezen kívül például Drouine (2003) az automatikus validálást előszűrőnek használja, így a terminusok kézi ellenőrzésénél kevesebb dolga van az annotátoroknak (nem kell a feleslegesen kinyert terminusokat kézzel kihúzogatnia).

L'Homme (2004), aki kijelenti, hogy a franciában a tipikus terminusszerkezetek a következők:

főnév-melléknév (pl. *mémoire morte* 'ROM')

főnév-főnév (pl. *carte-mère* 'alaplap')

főnév-prepozíció-főnév (pl. *machine à calculer* 'számológép')

főnév-prepozíció-főnévi igenév (pl. *base de données* 'adatbázis')

Ezek inkább a francia nyelvre jellemzők, elég csak belátni, hogy a magyar nyelvben nincs is prepozíció, így ezen mintákat a magyar nyelvre nem is lehet alkalmazni. Ezzel szemben más problémát is felvet ez a nyelvészeti megközelítés: ha csak ezt használnánk terminológiakivonatolásra, sok olyan kifejezés is illeszkedne, amely valójában nem terminus. Az előbbi mintákra visszatérve, könnyedén rájöhetünk, hogy a *home address* 'lakcím' is két egymást követő főnév, vagy a *big problem* 'nagy probléma' is melléknév-főnév kombináció, mégsem terminus. Ez számítógépes nyelvészeti terminusokkal azt jelenti, hogy a mintaillesztés elvén működő alkalmazásoknak ugyan nagy a fedése (a legtöbb terminust megtalálja), de kicsi a pontossága (a terminuslistában sok a nem releváns elem). Ezért ennél a módszernél mindenképpen szűrni kell a listát valamilyen statisztikai vagy egyéb módszerrel.

5.3.1.1. Mintaillesztés reguláris kifejezéssel

5.3.1.1.1. Bevezetés a reguláris kifejezések használatába

A reguláris kifejezéssel történő mintaillesztés az egyik legismertebb és legelterjedtebb mód. A reguláris kifejezés egy olyan szabványos – általában helyettesítő karaktereket is tartalmazó – karaktersorozat, amellyel leírható és felismerhető karaktersorozatok egy halmaza.[18]

A továbbiakban különböző példák segítségével mutatjuk be a legfőbb speciális karaktereket, amelyek sztringek ilyen típusú keresésére szolgálnak. Az első példa a francia hagyományos inflekciójú melléknevek keresésére szolgál. Ha a *bleu* 'kék' melléknév előfordulásait keressük a szövegben, akkor nemcsak a hímnem egyes számú *bleu* alakra kell keresni, hanem a nőnemű *bleue*, többes számú *bleus* és többes szám nőnemű *bleues* alakra is. Ennek megvalósítása különbözőképpen történhet, a legegyszerűbb eset a következő:

bleu|*bleue*|*bleue*|*bleues*

[18] A reguláris kifejezésekről bővebben Martín-Vide (2003)-ban olvashatunk.

A | jel jelentése „vagy”, tehát vagy az egyik, vagy a másik oldalán lévő kifejezést választja a mintaillesztés. Ez azonban egyszerűbben is megoldható, mert amennyiben több melléknevet is keresünk, akkor ezek felsorolása igencsak hosszadalmas, ezért egyszerűbb csak egy olyan kifejezést írni, amely azt írja csak elő, hogy a *bleu* mindenképpen szerepeljen a sztringben, utána pedig állhat *-e* vagy *-s*, vagy mindkettő, de csak ebben a sorrendben.

bleu(e)?(s)?[19]

A *?* azt jelenti, hogy az előtte álló, zárójelben lévő kifejezés vagy nullaszor, vagy egyszer fordulhat elő. Ez alapján, ha a *vert* ’zöld’ szót és annak összes előfordulását is keressük, akkor szintén két lehetőségünk van: a korábbi felsorolásos módszer, valamint az egyszerűsített változat.

bleu|bleue|bleue|bleues|vert|verte|verts|vertes

(bleu(e)?(s)?)|(vert(e)?(s)?)

(bleu|vert)(e)?(s)?

A + és a * jelekre természetes nyelvű szövegekben ritkán van szükségünk, ezért erre egy elvontabb példát hozunk. Ha olyan karaktersorozatot keresünk, amely a *bla*, *blabla*, *blablabla* stb. sorozatokra keres rá (elvileg mindhárom előfordulhat spontán szövegben a hitetlenség kifejezésére), akkor azt az alábbi reguláris kifejezés írja le jól, amelyben a + jel legalább egy előfordulást jelent:

(bla)+

A * bármennyi előfordulást jelent, így az alábbi keresőkifejezés bármely olyan karaktersorozatot felismer, amelynek kezdete *ve*, és bármennyi (akár 0) *t* követi (mint a *ve, vet, vett, vettt* stb. szavakban):

*ve(t)**

5.3.1.1.2. Reguláris kifejezések használata terminológiakivonatoláshoz

A mintaillesztés a TE szempontjából nagyon hasznos, hiszen tudjuk, hogy a terminusok nagy része valamilyen mintát követ, azaz vannak olyan speciális belső szerkezetek, amelyekre ezek illeszkednek. A mintát most nem a terminális szimbólumokon futtatjuk le, hanem nemterminális elemeken, amelyek a TE esetében morfoszintaktikai címkék, amelyek a szöveg szavairól megmondják, hogy azok milyen szófajúak.

[19] A zárójelezés a példákban sok helyen elhagyható, de a könnyebb átláthatóság kedvéért mindenhol kitettük.

A francia nyelvben vannak olyan szerkezetek, amik tipikusan terminusjelölt szerkezetek. Például három, egymás után álló főnév szinte biztos, hogy terminus, de két egymás után álló főnév is majdnem biztosan az, feltéve, hogy szakszövegben fordul elő (Nagy 2009a). Mivel a terminus lehet egyszavas főnévi kifejezés is, így az alábbi minta elég sok terminust kinyerhet:

N+

Ha a terminus N P N szerkezetű, például *base de données* 'adatbázis', akkor ezt már az előző minta nem ismeri fel. De a terminusok szerkezete alapján az N P N minta esetében a P N szinte bármennyiszer ismétlődhet, ezért erre az alábbi mintát javasoljuk:

N (P N)+

Így már nemcsak a *base de données* karaktersorozatot ismeri fel a gép, hanem a *système de gestion de base de données* 'adatbáziskezelő-rendszer' sztringet is, mert annak mintája N P N P N P N. Ha tovább vizsgáljuk a szintaktikai szerkezeteket, belátható, hogy ezen szerkezetbe bármikor bekerülhet egy melléknév, legtöbbször a végére: *système de gestion de base de données génériques* 'generikusadatbázis-kezelő rendszer', ezért a mintát tovább kell bővítenünk. Ekkor a következő reguláris kifejezéshez juthatunk:

N (P N)+ (A)?

A felsorolt mintákat ezután célszerű összevonni, amelyből egy újabb reguláris kifejezést kapunk:

N (P N)+ (A)?|N+

5.3.1.2. Mintaillesztés véges állapotú automatával

A számítástudományból jól ismert véges állapotú automaták (*finite state automata* vagy FSA) tekinthetők a legegyszerűbbnek az automaták közül: ez egy olyan gép, amely nem rendelkezik semmilyen tárral, és írási műveletet sem hajt végre. Csak végigmegy az input szövegen, és ha az automatába kódolt mintát illeszti a szövegre, akkor azt a szövegrészt megjelöli (Martín-Vide 2003).

A véges állapotú automatákat egy rendezett ötös $(Q, \Sigma, \delta, q_0, F)$ ír le:

„Az $M = (Q, \Sigma, \delta, q_0, F)$ rendszert nemdeterminisztikus automatának nevezzük, ha:

1. Q egy nem üres, véges halmaz, az *állapotok halmaza*,
2. Σ egy ábécé, az *input ábécé*,
3. $q_0 \in Q$ a *kezdő állapot*,
4. $P \subseteq Q$ a *végállapotok halmaza*,
5. $\delta : Q \times \Sigma \rightarrow \mathcal{P}(Q)$ egy leképezés, az *átmenetfüggvény*" (Fülöp 2004)

Az 1. pont az állapotok halmazát írja le: konvenció szerint az állapotokat q_0, q_1, q_2 … q_n

címkékkel jelöljük, ahol a q_0 a kezdőállapot (3. pont), és ezek közül tetszőleges számú végállapot is lehet, és ezen állapotokat az automatának tartalmaznia kell ($\subseteq$ Q, 4. pont). Az input ábécé a beolvasható elemeket tartalmazza: a szófaji címkék alapján történő TE esetén ezen szimbólumok közé sorolhatjuk például az *N* (főnév), *A* (melléknév) stb. szimbólumokat. Az 5. pontban az átmeneti szabályok találhatók, amelyek célja annak leírása, hogy egy adott állapotból egy adott input szimbólum beolvasásának hatására az automata mely állapot(ok)ba kerül. Az FSA akkor és csak akkor fogadja el a számára elemzésre adott input szót, ha a q_0 kezdőállapotból elindulva, a szó végigolvasása után egy p ($p \in P$) végállapotba kerül az 5. pontban leírt formájú átmeneti szabályok segítségével.

A véges állapotú automaták tervezésénél még két szempontot kell figyelembe venni: a determinisztikusságot, valamint a teljességet.

> „Egy $M = (Q, \Sigma, \delta, q_0, F)$ automata determinisztikus, ha teljesül, hogy minden $q \in Q$ és $a \in \Sigma$ esetén a $\delta(q,a)$ halmaz legfeljebb egyelemű." (Fülöp 2004)

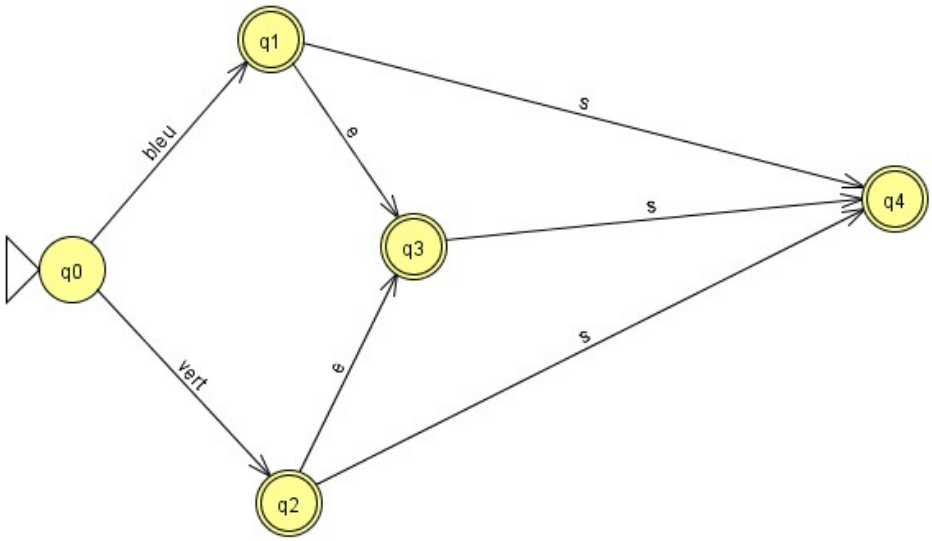
A determinisztikusság tehát azt jelenti, hogy az automata akármelyik állapotban is van, mindig csak legfeljebb egy állapotba mehet tovább (ami lehet ugyanaz is). A teljesség (más néven a teljesen definiáltság) definíciója a következő:

> „Egy $M = (Q, \Sigma, \delta, q_0, F)$ automata teljes vagy teljesen definiált, ha teljesül, hogy minden $q \in Q$ és $a \in \Sigma$ esetén a $\delta(q,a)$ halmaz legalább egyelemű." (Fülöp 2004)

Ez azt jelenti, hogy az automata mindig kerül valamelyik állapotba, akármelyik betűt is olvassa egy adott állapotban.

A véges állapotú automaták azon szavakat ismerik fel, amelyek reguláris kifejezéssel megadhatók: ezt tudjuk, mivel a számítástudomány egyik jól ismert tétele szerint a reguláris kifejezéssel megadható nyelvek osztálya megegyezik azon nyelvek osztályával, amelyek véges állapotú automatával felismerhetők. Ez alapján minden reguláris kifejezésre teljesül az, hogy készíthető belőle ugyanazt a nyelvet felismerő véges állapotú automata, és ez fordítva is igaz (Fülöp 2004).

Ennek illusztrálására a *(bleu|vert)(e)?(s)?* reguláris kifejezéshez egyszerűen megadhatunk véges állapotú automatát, amelyet a következő ábra szemléltet:



5.1. ábra. A *(bleu|vert)(e)?(s)?* reguláris kifejezéshez felírható FSA

A jelölésrendszer alapján a háromszöggel jelölt kör a kezdőállapotot, a szegélyes körök a végállapoto(ka)t jelölik. Egy karaktersorozatot akkor fogad el egy automata, ha a kezdőállapotból kiindulva valamely végállapotba kerül a gép. Az első automata a *bleu* bemeneti szóra a *q0* állapotból a *q1* állapotba kerül, amely végállapot, tehát azt az automata felismeri. A *vertes* szó esetén az automata a *q0* állapotból a *vert* karaktersorozatra a *q2* állapotba ér, majd az *e* betű olvasásakor a *q3* állapotba megy tovább, és onnan az *s* betű beolvasásakor a *q4*-be kerül, amely végállapot, tehát ezt a szót is felismerte. A *rix* szót a gép elutasítja, mert az input olvasásakor az automata nem végállapotban megáll.

Ekkor viszont az a kérdés merül fel, hogy miért nem mindegy, hogy a mintát reguláris kifejezéssel vagy automatával illesztjük-e, ha a kettőnek ugyanaz az eredménye. A hasonlítást mi magunk is el tudjuk végezni, hiszen korábbi tanulmányainkban már mindkettővel foglalkoztunk: a terminológiakivonatoláshoz reguláris kifejezéseket használtunk (Nagy 2008 és Nagy 2009a), magyar főnévi csoportok kinyerésére pedig véges állapotú automatát (Nagy 2009b).

Mivel a reguláris kifejezésekkel és a véges állapotú automatákkal felismerhető nyelvek osztálya megegyezik, így a terminuskinyerés során bármelyiket használhatjuk. Tapasztalataink szerint a reguláris kifejezéssel történő mintaillesztés kevésbé átlátható, ezért az automatás megoldás mellett döntöttünk. Minél több mintát fedezünk fel, amelyekkel terminusokat lehet felismerni, annál bonyolultabb lesz a reguláris kifejezésünk, ha azzal oldanánk meg. Ez semmiképp sem jó, mert ha már így is hosszú egy reguláris kifejezés, akkor már nehéz átlátni, hogy a mintát hogyan bővítsük. Automatába pedig egy

újabb állapot egyszerű felvételével bővíthetjük a mintát. Ezért a mintaillesztés esetében mindenképpen egy véges állapotú automata lehet – számunkra – átláthatóbb.

5.3.2. Újraíró szabályok terminusvariánsok kinyerésére

A terminusok egyik jellemzője, amely megkönnyíti annak számítógépes felismerését, az az, hogy a szövegben mindig ugyanúgy fordulnak elő. Ezzel ellentétben állnak a köznyelvi szavakkal, amelyeket nem kell feltétlenül ugyanúgy megismételni a szövegben, hogy arra visszautaljunk: ehelyett használhatunk szinonimákat, hiperonimákat vagy hiponimákat, vagy más kifejezéseket. Azonban itt is vannak kivételek, bár ezek száma és gyakorisága változó lehet, mégis figyelembe kell vennie minden terminuskinyerő rendszernek.

Daille (2005) szerint négy különböző variánsa lehet egy terminusnak: grafikai, inflekcionális, gyenge szintaktikai és morfoszintaktikai. A grafikai változat azt jelenti, hogy ugyanazt a terminust többféleképpen írják. Erre legjobb példa Boulaknadel és mtsai (2008), akik az arab nyelvre dolgoztak ki terminológiakivonatolót. Ebben a nyelvben a *p* és *h* karakterek gyakran egymással felcserélhetők, így a 'talajszennyezés' fogalmat az arabban lehet írni *tlwv Altrb**p*** vagy *tlwv Altrb**h*** kifejezéssel is. A francia nyelvben erre kevés példát találunk, így ezekkel nem foglalkozunk.

Az inflekcionális variánsok azt jelentik, hogy ugyanaz a terminus egy adott szövegben többféle inflekcióval is szerepelhet. Ez különösen az agglutináló nyelvek, például a magyar esetében igaz, de az indoeurópai nyelvek kevésbé jelentős ragozási rendszere sem elhanyagolható. Lényeges, hogy ugyanazt a terminust más ragozással ne vegyük fel egy terminológia adatbázisba, és gyakoriság vizsgálata esetén ne vegyük fel új terminusként. Ez a probléma könnyedén áthidalható egy lemmatizáló alkalmazás segítségével, amely csak az adott token szótövét veszi figyelembe: így például az *operációs rendszert, operációs rendszernek, operációs rendszertől* kifejezések egy terminus különböző variánsai, mégpedig az *operációs rendszer* terminusé. Mivel az alkalmazásunkban használunk szótövesítő alkalmazást, így ezen változatok külön vizsgálatára nem lesz szükségünk.

Nagy problémát jelent a morfoszintaktikai variánsok kezelése, amelyek olyan variánsok, amelyekben a terminus egy-egy eleme akár más szófajú is lehet. Erre a francia nyelvben gyakori eset lehet a prepozíció+főnév és melléknév váltakozás, amely azt jelenti, hogy a prepozíció utáni főnevet kicserélhetjük egy abból képzett, ugyanolyan jelentésű melléknévvel és a prepozíció elhagyásával. L'Homme (2004) példája erre a jelenségre:

épanchement de sang → *épanchement sanguin* 'vérömleny'. Mivel ezen terminusváltozatok egyenértékűként való kinyerése (tehát az, hogy a két változat egy terminusnak számítson) nagyon sok plusz munkát jelentene, ezért mi első megközelítésben ezen változatokat külön terminusjelöltnek vesszük. Ahhoz, hogy egyenértékű terminusvariánsnak tekintsük őket, szükség lenne nemcsak egy lemmatizáló alkalmazásra, hanem egy teljes tövesítést, azaz *chunking*ot végrehajtó alkalmazásra is, mert így az *épanche* és *sang* szópár kerülne a terminusjelöltek közé: viszont így a minták kinyerésére nem lehet szófaji címkéket használni. Valamint azt egyelőre nem tekintjük problémának, ha ezen változatok külön terminusként szerepelnének a végleges változatban, hiszen előfordulhat hogy ilyen esetekben van különbség a kettő jelentése között.

A legtöbb problémát tulajdonképpen a harmadik típusú variációk, azaz a gyenge szintaktikai variánsok adják. Jacquemin (2001) célja annak feltérképezése volt, milyen ilyen típusú változatok léteznek. Ezen variánsok feltérképezésére rengeteg metaszabályt alkalmaz, amelyek egyszerű újraíró szabályok. Többféle terminusváltozat létezik, leggyakoribbak a mellérendelő és alárendelő szókapcsolatok. A mellérendelés egyik metaszabálya a következő:

Metarule Coor(X1 → X2 X3) = X2 C4 X5 X3

E metakifejezés alapján a C4 egy mellérendelést kifejező kötőszó (*conjunction*), az X egy olyan nyelvtani kategória, amely a terminusokban gyakran szerepel, mondjuk N. Ez a szabály azért hasznos, mert ha például van két hasonló terminus, például *serum albumin* (savófehérje) vagy *egg albumin* (tojásfehérje), akkor az *egg and serum albumin* (tojás- és savófehérje) kifejezést a fent említett két terminusra bontja a metakifejezés (Jacquemin 2001).

A terminusvariánsok kezelése általában csak nyelvfüggő és szabály alapú alkalmazásként valósítható meg. A Jacquemin-féle (2001) terminusvariáns-kezelés is ezt mutatja: az adott nyelv koordinációs és egyéb szóösszetétel-kezelő alkalmazásai is szabály alapúak. A szakirodalomban ezzel kapcsolatban nem található olyan alkalmazás, amely ezt statisztikai módszerrel oldotta volna meg, bár tény, hogy a szabályok alapján nyert terminusokat lehet statisztikai módszerrel javítani.

5.3.3. Konnektívumok szűrése

Egy szöveg nemcsak szavak egymásutániságából áll, szükség van olyan elemekre is, amelyek biztosítják annak kohézióját, mégpedig úgy, hogy fenntartják az értelmezéshez

szükséges referenciális elemek lineáris kapcsolatát, egymásba ágyazódottságát (Riegel és mtsai 2009). Mivel a TE-t szövegen végezzük, nemcsak szavak egyfajta halmazából próbáljuk azokat kinyerni, ezért fontos, hogy figyelembe vegyük, hogy egy szöveg olyan elemeket is tartalmazhat, amelyek leginkább csak annak az egységei közötti kapcsolatot biztosítják. Többek között ilyenek a szövegkohéziót biztosító elemek, például a *más szóval, például* vagy *következésképpen*. A TE során ezért lesznek olyan szókapcsolatok, amelyek a terminusoknak nem részei, és ezért ezeket ki kell szűrni.

A szöveg kohéziós elemei közül a logikai konnektívumokat fogjuk csak figyelembe venni, de ezen kívül másfajta ilyen elemek is léteznek. A tematikus progresszió azért fontos, mert egy szöveg attól koherens, hogy időnként visszautal egy már kijelentett tartalomra vagy az olvasó egy előfeltételezett tudására. Egy szövegbeli mondat tehát általában két részre oszlik: (1) egy már említett vagy ismert dologról (2) mond valami újat. Az ismert részt hívjuk témának, az új információt rémának. A szóismétlés elkerülése és a szöveges kohézió fenntartása végett használjuk még az anafórákat, amelyek részleges vagy teljes egyezéssel visszautalnak egy már korábban említett elemre, ezzel is megadva a szöveg fonalát (Riegel és mtsai 2009).

A konnektívumok a szöveg strukturálását és az elemei közötti kapcsolatot biztosítják: jelzik azt, hogy milyen típusú viszony áll fenn vagy a tagmondatok vagy egyéb szósorozatok között. Egy szöveg elemeit elkülönítik egymástól, vagy épp ellenkezőleg, közelebb hozzák őket egymáshoz, ezzel kiegészítve az írásjelek biztosította lehetőségeket (Riegel és mtsai 2009).

Schneuwly és mtsai (1989) szerint a konnektívumok három fő tulajdonsággal rendelkeznek: (1) a tagmondatoknak nem nélkülözhetetlen elemei, (2) a tagmondatokat egymáshoz kapcsolják vagy sorozatba rendezik a szövegkörnyezetbe történő helyezésükkel, (3) nincsenek – szám- vagy személybéli – egyeztetés alá vonva az előtte és utána álló elemekkel. Ez utóbbi különíti el őket az anafóráktól, amelyeket számban és személyben egyeztetünk az antecedensükkel.

Riegel és mtsai (2009) szerint a konnektívumoknak három fő fajtája létezik: (1) szövegstruktúráló elemek, (2) a kijelentés forrásának jelzői, (3) argumentatív konnektívumok[20]. A szövegstrukturáló elemek időben és térben helyezik el a szöveg által leírt tartalmat, ezért általában narratív szövegekben fordulnak elő. Ilyen elemek az *alors* 'akkor', *ensuite* 'aztán', *plus loin* 'távolabb' vagy *à droite* 'jobbra', amelyek közül az első

[20] Saját fordítások. Francia megfelelőik: *organisateurs textuels, marqueurs de prise en charge énonciative, connecteurs argumentatifs*.

kettő időbeli, a második kettő térbeli pontosítást tartalmaz. A második típusú konnektívum feladata az, hogy a szövegben azt fejezze ki, hogy az azt megelőző vagy követő rész kinek a nézőpontját tükrözi. Erre példa *selon X* 'X szerint', *pour X* 'X számára'. Riegel és mtsai (2009) ide sorolja még az olyan szövegszervezési konnektívumokat is, mint *c'est-à-dire* 'azaz', *en d'autres termes* 'más szóval', *autrement dit* 'más szóval' vagy az olyan záró formulákat mint *en somme*, *en fin de compte* 'összegzésképp'. A harmadik típusú kategóriába tartozik az *en effet* 'ugyanis', *certes* 'bizonyára', *par exemple* 'például' vagy az *or* 'továbbá'.

A TE során tehát fontos, hogy elkülönítsük azokat az elemeket, amelyek a szaknyelvhez köthetők (terminusok) azoktól, amelyek csak a szöveg kohézióját biztosítják (konnektívumok), mert ez utóbbiak biztosan nem lehetnek terminusok, vagy nem lehetnek terminusok részei.

Mivel ezen elemek halmaza zárt, így ezeket könnyebb szabály alapú módszerekkel kiszűrni: legegyszerűbb módszer ezek szó szerinti keresése, ezért is tettük a konnektívumok szűrését a szabály alapú modulok közé.

5.3.4. Terminológiai helyzet

Kis Á. (2007) szerint a terminológiai helyzet egy olyan rész a szakmai szövegben, amelyet úgy lehet definiálni, mint egy olyan szövegkomponens, amit az olvasó másképp érzékel, mint a többi részt: az értelmezéséhez az adott kontextushoz képest többet kell látnia, mert ezekben a helyzetekben egy terminusnak kell megjelennie. A terminológiai helyzetet lehet űrként is értelmezni, amelybe csak terminus kerülhet. Ennek az a következménye, hogy a terminológiai helyzetben szereplő szövegegységet az azt olvasó terminusnak véli, sőt, akármi is az, el is fogadja annak.

Az általánosabb szakmai szövegeknek (például a didaktikusabb, magyarázó jellegű leírásoknak) megvan az az előnye, hogy abban az új fogalmak valamilyen módon definiálva is vannak. Ebből az következik, hogy ha megtaláljuk azokat a szerkezeteket, amelyek tipikusan jellemzőek a definíciókra, akkor megtalálhatjuk a hozzájuk tartozó terminusokat is.

A magyar nyelvben például Kis Á. (2007) szerint a terminust és a hozzá tartozó definíciót tartalmazó tipikus definíciós környezetek – többek között – a következők:

(1) megnevező mondat: ide tartoznak például a ...-*nak nevezzük* típusú szerkezetek, ahol az azt megelőző elem a definiálandó terminus (a példában a *biometrikus adat*), ami

pedig azt követi, az a definíció (példánkban az *embereket testi jellemzőik alapján azonosítani képes adatok*).

> *Biometrikus adatoknak nevezzük az embereket testi jellemzőik alapján azonosítani képes adatokat.*

(2) névszói állítmány, ahol az alany a definiált terminus (példánkban a *disszertáció*), a névszói állítmány pedig annak a definíciója (példánkban a *tudományos mű, amit szerzője a doktori fokozat megszerzéséhez készít*).

> *A disszertáció egy tudományos mű, amit szerzője a doktori fokozat megszerzéséhez készít.*

(3) egyéb, összetett szerkezetek, például a kijelölő jelző (példánkban a *mint például kézerezet, írisz, ujjlenyomat*), amely az előtte álló terminus (példánkban a *biometrikus adat*) definícióját tartalmazza.

> *A biometria az emberek géppel is kezelhető, egyedi jellemzőjét (ún. biometrikus adatait) – mint például kézerezet, írisz, ujjlenyomat – alkalmazza az adott személy beazonosítására.*

5.4. A terminológiakivonatolók statisztikai moduljai

E fejezetben bemutatjuk, melyek azok a statisztikai módszerek, amelyek A TE-ben alkalmazhatók. Mint ahogy igaz volt a szabály alapú modulokra, úgy itt is érvényes, hogy a statisztikai elemek mind a terminusjelölt-lista első összeállításakor, mind a szűrés folyamatában is hasznosak lehetnek.

A statisztikai elemek általában két csoportra oszthatók: vannak olyanok, amelyek *termhood*- és vannak olyanok, amelyek *unithood*-mértékeket mérnek. Az utóbbi azt mutatja meg, hogy egy adott többszavas terminus elemei mennyire tartoznak össze, azaz mennyire erős a kohézió a terminusok különböző elemei között. A *termhood*-mérték, amit már egyszavas terminusokra is alkalmazhatunk, azt mondja meg, hogy egy kinyert terminusjelölt mennyire köthető egy adott szakterülethez. Ha ez az érték alacsony, az a kifejezés általános nyelvi korpuszban vagy más szakterületen is előfordulhat, így kisebb eséllyel terminus (Wong és mtsai 2008).

A legtöbb TE-alkalmazás ezek kombinációját foglalja magában; ezen mértékek együttes használata nagyban segít abban, hogy csak a terminusokat nyerjük ki egy adott szövegből. Ha a terminusjelölt *termhood*-mértéke nagy, a terminus tényleg egy adott tudományterülethez tartozik, és ha a *unithood*-értéke is magas, akkor a terminusjelölt

egésze terminus, azaz sem az őt tartalmazó nagyobb szócsoport, sem a terminusjelölt része nem az.

5.4.1. Gyakoriság vizsgálata – *termhood*-mértékek

A TE általános tendenciája, hogy a gyakran előforduló szócsoportokat kezelik az alkalmazások, és azokat veszik terminusjelöltnek. Ennek nagyon sok hátránya van, és nagyban függ annak a korpusznak a méretétől, amelyből a terminusokat próbáljuk kinyerni. Mindamellett azt a tényt elfogadhatjuk, hogy ha nagy terjedelmű a korpusz, akkor ugyanaz a terminus valószínűleg gyakran ismétlődik.

Azonban a gyakoriság vizsgálata nem elegendő, mert nagyon sok kifejezés csak egyszer vagy kétszer fordul elő egy adott korpuszban – főleg ha a korpusz mérete nem is túl jelentős (Boulaknadel és mtsai 2008; Yang és mtsai 2008; Cabré és mtsai 2001).

Ezen kívül egyéb hátránya is van, ha csak a gyakoriságot, azaz egy adott terminus adott szövegbeli előfordulási arányát, vesszük alapul. Léteznek ugyanis olyan kifejezések, szavak, amelyek egy adott nyelvben, akár az általános, akár a szaknyelvben, sokszor szerepelnek, mégsem lehetnek terminusok. Ezek között a főnevek vagy főnévi csoportok azok, amelyek érdekesek, mert a legtöbb terminológiakivonatoló inkább ezekre koncentrál. Például az angol nyelvben a *fact* 'tény', *machine* 'gép' stb. önmagukban gyakran fordulhatnak elő bármilyen korpuszban, azonban ezeket biztosan nem kell terminusoknak vennünk, esetleg csak a belőlük képzett összetett terminusokat, például *vending machine* 'árusító automata'.

A gyakran előforduló kifejezések szűrésére leggyakrabban használt technika egy referenciakorpusz használata. A referenciakorpusz egy olyan szöveggyűjtemény, amely nem szaknyelvi, hanem általános nyelvi szövegeket tartalmaz. Ezek legtöbbször az írott sajtó területéről jönnek: ha újságcikkeket választunk, ügyelnünk kell arra, hogy valamilyen általános rovatból jöjjön. A referenciakorpusz lényege tulajdonképpen abban áll, hogy ha a szaknyelvi korpuszban terminusgyanús kifejezésre bukkanunk, akkor arról könnyedén eldönthetjük, hogy tényleg terminus-e, ha megvizsgáljuk annak gyakoriságát vagy puszta jelenlétét a referenciakorpuszban. Ha a szaknyelvi korpuszban egy szóegyüttes gyakran szerepel, de a referenciakorpuszban alig vagy egyáltalán nem (azaz a szaknyelvben nagyobb arányban fordul elő), akkor valószínűsíthető, hogy az terminus. Referenciakorpuszt majdnem minden terminológiakivonatoló használ, mint például Drouin (2003), Yirong és mtsai (2006).

5.4.1.1. TF-IDF

A termhood típusú statisztikai modulok közül a tf-idf metrika az, amit alapnak tekintünk. A tf-idf a *term frequence – inverse document frequence* rövidítése, és arra használható, hogy megkapjuk, milyen valószínűséggel szerepelhet egy terminus egy adott dokumentumban vagy dokumentumcsoportban, így ezáltal azt vizsgálhatjuk, hogy egy terminus mennyire kötődik egy adott dokumentumtípushoz, vagyis szakterülethez.

A TF-IDF mérték kiszámítása a következő[21]:

$$w_{(t,d)} = tf_{(t,d)} \cdot \log\left(\frac{N}{df_t}\right)$$

A képletben $w_{(t,d)}$ egy olyan érték, amely azt mutatja meg, hogy a *t* kifejezés a *d* dokumentumhoz mennyire kapcsolódik, a $tf_{(t,d)}$ a *t* kifejezés *d* dokumentumbéli előfordulásának számát jelzi, *N* a dokumentumok száma, df_t pedig azt adja meg, hogy a *t* ezekből hány dokumentumban szerepel. A képletből az következik, hogy ha egy *t* kifejezés csak egy dokumentumban fordul elő, akkor ahhoz a dokumentumhoz nagy értéket fog kapni, ha pedig egy *t* kifejezés minden dokumentumban nagyjából ugyanannyiszor szerepel, akkor ez az érték sokkal kisebb lesz. Ha egy *t* kifejezés minden dokumentumban szerepel, akkor a törtnek mind a számlálójában, mind a nevezőjében ugyanaz az érték szerepel, így az összérték 0 (mert log 1 = 0), de akkor is 0 ez az érték, ha az adott dokumentumban az a szó nem is szerepel.

Ezzel a módszerrel akkor lehet probléma, ha egy adott területhez több dokumentum is tartozik: ekkor ugyanis ugyanahhoz a kifejezéshez minden dokumentum esetén különböző érték tartozhat, és mivel azt szeretnénk megtudni, hogy egy adott terminus egy szakterülethez köthető-e, így ezen értékeket egységesíteni kell. Ekkor viszont döntenünk kell, hogyan egységesítsük ezeket az értékeket, hiszen ha csak azt vizsgáljuk, hogy egy terminus mennyire köthető egy szakterülethez, akkor arra már nem vagyunk kíváncsiak, hogy külön-külön az egyes dokumentumokban ez a terminus mennyire gyakran szerepel. Ha átlagoljuk a különböző értékeket, akkor lehet, hogy alacsony értéket kapunk az olyan esetekben, ahol csak egy-két dokumentumban szerepel az adott terminus, még ha ott sokszor is. Választhatjuk azt is, hogy ha egy szakterületi dokumentumban nagy a tf-idf érték, akkor azt mindenképpen elfogadjuk.

Ezen probléma egyik megoldására tett kísérletek egyikének tekinthető a Basili és mtsai (2001) által kidolgozott CW (*Contrastive Weight*, azaz kontrasztív súly) érték, amely

[21] A TF-IDF metrika alapjairól bővebben l. Manning és mtsai (2008), terminológiakivonatoláshoz kapcsolódó aspektusairól l. Salton és Backley (1988) vagy Evans és Lefferts (1995).

már nem azt nézi, hogy egy *t* kifejezés egy *d* dokumentumra releváns-e, hanem azt, hogy egy kifejezés mennyire releváns egy szakterületre. A CW kiszámítási módszere azt sugallja, hogy az általános nyelvi kifejezések eloszlása hasonló minden szövegtípusban, de a terminusok csak egy adott dokumentumtípusban jelennek meg. Egy adott *a* egyszavas kifejezés kötődése egy *d* szakterülethez a következő képlettel írható le:

$$CW(a) = \log f_{ad} \left(\log \frac{\sum_i \sum_j f_{ij}}{\sum_j f_{aj}} \right)$$

Az f_{ad} az *a* terminusjelölt előfordulásának száma a *d* szakterületen, a tört számlálója azt írja le, hogy az összes területen mennyi a terminusjelöltek száma, a nevező pedig azt írja le, hogy az *a* terminusjelölt hányszor szerepel összesen az összes tudományterületen.

A többszavas kifejezések esetén ennek egy módosított változatát használják, amire szerintük azért van szükség, mert a komplex terminusok ritkábban fordulnak elő, és szeretnék, ha a puszta gyakoriságukkal nem kerülnének hátrányba. Ilyenkor az összehasonlítás alapjaként a komplex főnévi csoport fejét vizsgálják, így a képlet a következőre módosul:

$$CW(a) = f_{ad} CW(a^h)$$

Az f_{ad} az összetett terminusjelölt előfordulásának száma a *d* szakterületen, az a^h pedig az a szócsoport fejét jelképezi. Ezáltal amikor azt keressük, hogy az egész korpuszban hányszor fordul elő az összetett terminus, akkor csak azt keressük, hogy a feje hánysor szerepel a korpuszban. Ennek az a legfőbb hátránya, hogy nem mindig dönthető el egyértelműen, hogy mi a fej, mert ha például a *contact lense* összetett terminus gyakoriságát nézzük, akkor nem elég csak a *lense* vagy csak a *contact* előfordulásait vizsgálni, ott a kettő együttesére kell rákeresni. Az félrevezető lehet, hogy ha a korábban említett *soft contact lense* kifejezést tekintjük terminusnak, mert akkor ha a *contact lense*-t keressük meg a korpuszokban, akkor az nem mondja meg, hogy a *soft* kifejezéssel ez együtt áll vagy sem. Természetesen ilyenkor újabb szűrési algoritmussal ezen változtathatunk.

Egy szintén a TF-IDF algoritmusra épülő módszer a Yirong és mtsai (2006) által kidolgozott módszer, az úgynevezett DCA (*Document Crossing Algorithm*), amely ennek egy egyszerűsített változata. A kiszámítandó érték neve itt *Termhood(w)*, ahol *w* az adott terminusjelölt:

$$Termhood(w) = 1 - \frac{DF(w)}{f(w)}$$

A *DF(w)* azon dokumentumok száma, amely *w*-t tartalmazza, *f(w)* pedig a *w* összes előfordulásának száma. Ez az érték mindig 1 és 0 közötti érték, és akkor közelít az egyhez, ha a tört értéke minél közelebb van a nullához. Ez az érték akkor nagy, ha a terminusjelölt sokszor fordul elő nagyon kis számú dokumentumban, és akkor nulla, ha például minden dokumentumban pontosan egyszer szerepel. Azonban láthatjuk, hogy ennek a módszernek elég sok hátránya van: ha nincs referenciakorpusz, akkor a minden dokumentumban szereplő terminusokat biztos kiszűri, valamint nem tesz különbséget aközött, hogy bizonyos dokumentumokban hányszor szerepel az adott terminusjelölt, az előfordulásokat egyszerűen átlagolja.

Yirong és mtsai (2006) kibővítette a DCA-t, ami sokkal jobb eredményekhez vezetett, hiszen azt mutatta meg, hogy egy terminus mennyire köthető egy adott szakterülethez. Ennek a neve DRA (*Domain Relativity Algorithm*), és az alábbi formula írja le ennek kiszámítását:

$$Association(d,w) = \frac{p(d,w) - p(d)p(w)}{p(d,w)}$$

A *w* egy terminusjelölt, *d* egy tudományterület, *p(d,w)* annak valószínűsége, hogy egy *w* terminus egy *d* területhez tartozik, *p(d)* azon dokumentumok aránya, amelyek az adott területhez kötődnek, *p(w)* pedig *w* előfordulási valószínűsége az egész korpuszban. A *p(d,w)* érték úgy számolható ki, hogy megnézzük, hogy milyen arányban szerepel az adott terminusjelölt az adott terület korpuszában, a *p(w)* pedig a korábban említett *DF(w)* és az összes dokumentum számának aránya. Yirong és mtsai (2006) szerint a tf-idf nem tudja kezelni a gyakran előforduló, de nem terminusértékű szerkezeteket, ezért is vezette be a DCA-t, amely annál sokkal jobb eredményt ért el.

A TF-IDF algoritmusnak még rengeteg változata létezik, például a KF-IDF (Kurz és Xu 2002), de ezeket nem tekintjük át bővebben, hiszen a működési elve hasonló a többi esetben is.

5.4.1.2. Log-likelihood

Másik igen elterjedt *termhood*-mérték a log-likelihood, amelyet gyakran csak LL-ként említenek. Ezen mértéket használja például Macken és mtsai (2008) vagy Cohen (1995). A *log-likelihood* jelen részletes leírásakor Rayson és Garside (2000) tanulmányára hivatkozunk, amely egy igen áttekinthető útmutató ezen algoritmus feltérképezésére. Ez utóbbi tanulmány célja egyébként nem közvetlenül a TE, hanem korpuszok összehasonlítása az abban található szavak alapján.

A *log-likelihood* értéket mindig egy adott szóra kell megvizsgálni. Először is ki kell számolni az adott szó várt értékét (*E*) minden egyes korpuszban: a továbbiakban az *i* paraméter a dokumentum számára hivatkozik. Így E_i értéke egy adott szóra illetve adott dokumentumra a következő:

$$E_i = \frac{N_i \sum_j O_j}{\sum_i N_i}$$

Az N_i az *i*-dik dokumentum szavainak száma, O_i pedig az adott szó előfordulásainak száma az *i*-dik szövegben. A várt értékek figyelembe veszik a korpusz nagyságát, így azokat nem kell normalizálni. Eztán kiszámítható az adott szó *log-likelihood* értéke, amit most LL-lel jelölünk:

$$LL = -2 \ln \lambda = 2 \sum_i O_i \ln \left(\frac{O_i}{E_i} \right)$$

Macken és mtsai (2008) kimutatta, hogy az alacsony LL-értékek azt mutatják, hogy az adott szó sokszor előfordul más korpuszokban is, így az biztosan nem terminus, a magas érték pedig azt jelenti, hogy bizonyos dokumentumokban szerepel csak gyakran, így az valószínűbb, hogy terminus.

5.4.1.3. *Weirdness*

Ahmad és mtsai (1999) nevéhez fűződik a *weirdness* nevű *termhood*-érték, amely eredetileg nem TE-célokra készült. Az általuk kifejlesztett WILDER alkalmazás célja egy olyan dokumentumosztályozó és -kinyerő rendszer, amely gyakoriságvizsgálatra és korpuszösszehasonlítási technikákra épít. Ehhez terminusokat nyer ki, amelyek alapján behatárolja a szöveg témáját.

A *weirdness*-ráta kiszámítása egy nagyon egyszerű képlettel írható le:

$$weirdness(w) = \frac{\frac{f_s(w)}{t_s(w)}}{\frac{f_g(w)}{t_g(w)}}$$

A *w* a vizsgált szó, $f_s(w)$ a *w* előfordulási száma a szakszövegben, $f_g(w)$ a *w* előfordulási száma az általános korpuszban, $t_s(w)$ a szakszöveg összes tokeneinek a száma, míg $t_s(w)$ az általános korpusz tokeneinek száma. Azon szavak esetében, amelyek mind az általános, mind a szakszövegben gyakran és ugyanolyan aránnyal fordulnak elő, ott ezek értéke 1

körüli, míg a terminusoké, amelyek jobban meghatároznak egy szakszöveget, ennél sokkal nagyobb.

E formulával kapcsolatban csak egy probléma van: mi történik akkor, ha a terminus nem fordul elő általános nyelvi korpuszban? Akkor a fő tört nevezőjébe 0 kerül, mert $f_g(w)$ értéke is 0, ekkor viszont a 0-val való osztás miatt hibába ütközünk. Ilyenkor valamilyen normalizációs technikát is alkalmaznunk kell, ami további időigényes számításokhoz is vezethet.

5.4.2. *Unithood*-mértékek

Ebben az alfejezetben a különböző *unithood* mértékeket mutatjuk be. Ezek azok az értékek, amelyek arra használhatók, hogy megmutassuk, az adott terminusjelölt szegmensei mennyire tartoznak egybe. Ha ugyanis azt tapasztaljuk, hogy a terminusjelöltet valamely pontja előtt és után szétválaszthatjuk két külön terminusjelöltre, akkor azokat kell megtartani a terminusjelölt-listában (Wong és mtsai 2008).

5.4.2.1. Mutual Information

A legalapvetőbb és szinte leggyakrabban felhasznált *unithood*-mérték a *Mutual Information* (MI), (Church és Hanks 1990). Az MI mértékét az *a* és *b* szópárra vonatkozóan az alábbi módon számoljuk ki:

$$MI(a,b) = \log_2 \frac{p(a,b)}{p(a)p(b)}$$

A *p(a)* és *p(b)* rendre *a* és *b* előfordulási valószínűsége, *p(a,b)* pedig *a* és *b* együttes előfordulásának valószínűsége. Jól mutatja e módszer hatékonyságát L'Homme(2004), aki ezt két szópáron tesztelte: *air comprimé* ('sűrített levegő') és *air sont* ('levegő vannak', például olyan mondatban, hogy *Les conditionneurs d'air sont placés sous le plafond* 'A légkondicionálók a plafon alá vannak helyezve'). A vizsgált szövegben 50 000 szó volt, az *air* szó 123-szor (p(*air*)=0,00246), a *comprimé* szó 85-ször (p(*comprimé*)=0,0017), a *sont* szó 512-szer (p(*sont*)=0,01024) fordult elő. A vizsgált szóösszetételek közül az első 81-szer (p(*air,comprimé*)=0,00162), a második pedig 6-szor fordult elő (p(*air,sont*)=0,00004). Az IM-értékek rendre 8,6344 és 0,6671 lettek, amelyek közül az első látszólag is sokkal jelentősebb, tehát azok a szavak jobban összetartoznak, mint a másik kettő. Ha meghúzunk egy határt, amely felett elfogadjuk egy szópár asszociativitását, akkor végeredményként

mindig olyan párokat kapunk, amelyek biztosan erősen kollokálnak, tehát valószínűleg terminusok is.

5.4.2.2. C-érték

A C-érték, valamint az NC-érték, amelyeket gyakran használnak különböző TE-alkalmazások fejlesztésekor, pontosabb mutatója karaktersorozatok egybetartozásának. Ennek kidolgozói Frantzi és Ananiadou (1997), majd ezeket Frantzi és Ananiadou (1999), valamint Maynard és Ananiadou (2000) fejlesztették tovább.

A TE-hez már csak az NC-értéket használják, de ahhoz előbb szükség van a C-érték kiszámításra is. A C-érték elsősorban a gyakoriságra épít: miután megtörtént a terminusjelölt-lista összeállítása, utána alkalmazható ezen lista tagjaira az alábbi formula:

$$C\text{-}value = \log_2|a| \cdot f(a) \qquad ha\ a\ nem\ beágyazott$$

$$C\text{-}value = \log_2|a| \cdot f(a) - \frac{1}{P(T_a)} \sum_{b \in T_a} f(b) \qquad egyébként$$

Az *a* a vizsgált terminusjelölt, *f(a)* annak előfordulási száma, |a| a terminusjelölt hossza szavakban mérve, T_a azon terminusjelöltek halmaza, amely tartalmazza *a*-t, $P(T_a)$ pedig azon terminusok száma, melyek tartalmazzák *a*-t, és hosszabbak is nála. Úgy járunk el, hogy ha egy olyan terminust találunk, amelynek hossza maximális, akkor annak C-értéke tulajdonképpen annak hosszából és előfordulási gyakoriságából áll. Ha ez nem maximális terminus, akkor megnézzük azon terminusjelölteket, amelyek tartalmazzák *a*-t: ha van ilyen, megnézzük azok összes előfordulásának számát, majd átlagoljuk az összeget, és azt kivonjuk *f(a)* összes előfordulásából.

Ezen értéktől már egyenes út vezet az összesített NC-értékhez, amelynek kiszámítására több formula is létezik, de a szókörnyezet súlyértékét minden esetben figyelembe kell venni. Az első azt veszi figyelembe, hogy a C-érték sokkal pertinensebb jegy a szövegkörnyezetnél, így a C-érték 80, a szövegkörnyezet értéke pedig 20%-ban számít bele a végleges értékbe. A szövegkörnyezet értékének (*wei*) kiszámítását az 5.4.3. alfejezetben (*Kontextus figyelembe vétele*) írjuk le. Így az NC-érték az alábbi formulával írható le:

$$NC\text{-}value(a) = 0{,}8 * C\text{-}value(a) + 0{,}2 * wei(a)$$

Ennek a mértéknek egy másik változata pedig az alábbi formulával írható fel[22]:

[22] Itt hozzátennénnk, hogy Wong és mtsai (2008) szerint ebben a formulában a logN helyett csak N szerepel. Ezt azonban elvetettük, mert az 1/N érték sokkal kisebb lenne ezáltal, ha a korpusz nagy méretű, ha pedig sok a tört értékében a 0, akkor azzal nehezebb számolni. Például ha a korpusz 50000 szavas, akkor 1/N értéke 0,00002, míg 1/logN értéke kettes alapú logaritmusnál 0,0641.

$$NC\text{-}value(a) = \frac{1}{\log N} \cdot C\text{-}value(a) \cdot wei(a)$$

Az *a* a vizsgált terminusjelölt, *N* a korpusz mérete szavakban (a könnyebb számolhatóság végett csak annak logaritmusát vesszük), és a *wei(a)* a jelölt környezetében lévő szavak összesített súlyértéke.

Vu és mtsai (2008) szerint a C/NC érték jelenleg is a leginkább elterjedt terminológiakivonatolási módszer, amely még jobban alátámasztja az algoritmus hatékonyságát. Jóllehet ezt a metódust először csak az angolra fejlesztették ki, de már azóta más nyelveken is kipróbálták, például a szlovénen (Vintar 2004) vagy a japánon (Mima és Ananiadou 2001), és ráadásul több különböző szakterületen is, például az orvostudományban (Frantzi és mtsai 1998) vagy az informatikában (Milios és mtsai 2003).

5.4.2.3. Mutual Expectation

Annak eldöntésére, hogy egy kivonatolt terminusjelölt egésze vagy csak része kerüljön be egy terminológiai adatbázisba, gyakran használják a Mutual Expectation (ME) értéket, amelyet többek között Dias és Kaalep (2003), valamint Daille (1995) is használ. Ezen érték kiszámolásához először szükség van a normalizált várt érték (NE – *Normalized Expectation*) kiszámításához, amihez általában elengedhetetlen az összetartozó szócsoport kohéziójának vizsgálata. Tehát ha arra vagyunk kíváncsiak, hogy a Dias és Kaalep (2003) által is említett *take into custody* elemei mennyire tartoznak össze, akkor megvizsgáljuk ennek a szóhármasnak a valószínűségét, majd azt, hogy a *custody* milyen valószínűséggel jön egy *take into* szópár után, majd azt, hogy az *into* milyen gyakran köti össze a *take* és a *custody* szót, valamint azt, hogy milyen várható értékkel előzi meg a *take* az *into custody* kifejezést. Így jutunk az alábbi formulához, ahol az ő esetükben az n-gram az eredeti szóhármas, az n–1-gram pedig a szókettesek valószínűségét mutatja:

$$NE = \frac{prob(n\text{-}gram)}{\frac{1}{n}\sum prob(n-1\text{-}grams)}$$

A normalizált várt érték fő célja, hogy a kohézió vizsgálatán keresztül megadja annak a veszteségnek az értékét, amely abból fakadna, ha az egyik elemet kihagynánk. Tehát a kohézió egy összetartozó szócsoport esetén azt jelenti, hogy egy ilyen szócsoport minél inkább eltűri azt, hogy egy-egy eleme kiessen, akkor annál kevésbé koherens, így az NE értéke is alacsony lesz. Ha nagyon sokszor előfordulnak önmagukban is az eggyel

kevesebb elemet tartalmazó szókapcsolatok, annál nagyobb lesz azok átlaga, tehát a nevező is, és így annál kisebb lesz az NE értéke.

A végső ME-értéket a normalizált várható érték alapján számoljuk ki:

*ME = freq(n-gram)*NE(n-gram)*

5.4.2.4. IR/CR

A *unithood* mérésére egy eléggé egyedi és összetett módszert fejlesztettek ki Yirong és mtsai (2006). A *unithood* mérését két részre osztják: belső és külső. A külső mérés során azt nézik meg, hogy az adott terminusjelöltnek mi van a bal és jobb oldalán, hogy eldöntsék, az autonóm vagy egy nagyobb terminus része. A belső mérés pedig azt nézi meg, hogy a terminusjelölt milyen viszonyban áll az alsztringjeivel saját magán belül.

Először a külső mérés történik meg, mégpedig az LD (*left dependent rate*) és RD (*right dependent rate*) segítségével. Ezek formulái a következők:

$$LD(w) = \frac{\max_{a \in A} f(aw)}{f(w)}$$

$$RD(w) = \frac{\max_{b \in B} f(wb)}{f(w)}$$

Az *f(w)* a *w* terminusjelölt előfordulásának száma, *A* az összes olyan elem halmaza, amely megelőzi *w*-t, *B* azon elemek halmaza, amelyek jobbra állnak *w*-től. Ez tehát azt mutatja statisztikailag, hogy *w* mennyire függetlenül a szomszédos elemeitől: minél kevesebb különböző elem lehet a szomszédja, annál kisebbek ezek az értékek. Azonban nem elég csak az egyik értéket vennünk, mert mindkét szomszédra szükségünk van, ezért vesszük ezek egyfajta összesített értékét, amelyet Independent Rate-nek (IR) neveznek. Ennek értékét az alábbi formula írja le:

$$IR(w) = \left(1 - \frac{1}{f(w)}\right) * \sqrt{(1 - LD(w)) * (1 - RD(w))}$$

A külső mérés után következhet a belső mérés használata, amely során a karaktersorozat belső függetlenségét nézzük. Ezt az értéket Connectivity Rate-nek (CR) nevezik. Ehhez azt tételezzük fel először is, hogy *w* különböző szegmentumokkal rendelkezik, tehát *w* felírható $w_1w_2...w_n$ formátumban. A CR-rátát mindig két szomszédos elemre számolhatjuk ki, amelynek formulája a következő:

$$CR(w_i w_{i+1}) = \frac{p(w_i w_{i+1}) - p(w_i)p(w_{i+1})}{p(w_i w_{i+1})} * \sqrt{LD(w_i w_{i+1}) * RD(w_i w_{i+1})}$$

A *p(w)* a szokásos érték (*w* valószínűsége), $CR(w_iw_{i+1})$ a két vizsgált szomszédos elem konnektivitási rátája. Ha ez a $CR(w_iw_{i+1})$ nagyon alacsony, az azt jelenti, hogy a *w* terminus azon a helyen szétbontható, tehát ott esetleg két újabb terminusjelöltre választható szét. Természetesen sok helyen lehet alacsony ez az érték a *w* karaktersorozat esetében, ezért mindig a leggyengébb helyen kezdjük a szétbontást, tehát meg kell határoznunk *w* töréspontjainak minimumát, és ott kell tovább folytatnunk a szétbontást, ha szükséges. A minimum kiszámításához az alábbi formula áll rendelkezésünkre:

$$CR_{min}(w_1 \ldots w_n) = \min_{1 \le i \le n-1} (CR(w_i w_{i+1})$$

Eztán már csak egy lépés van hátra, az összesített *unithood*-érték kiszámítása, amely az alábbi módon történik:

$$Unithood(w) = IR(w) * CR_{min}(w)$$

Ezen egyesített értékhez beállítottak egy küszöbértéket, amely 0,0157. Ezzel a végleges korpuszban 72%-ot értek el, de a leggyakoribb 100 terminus esetében a pontosságuk 91% volt.

5.4.2.5. Egyéb mértékek

A fentebb felsoroltakon kívül még számos mérték létezik, amely mind azt hivatott eldönteni, hogy két elem mennyire kollokál egymással. Ezek hátránya nagyobbrészt abban áll, hogy csak két elemre alkalmazható, és csak azok kohézióját lehet mérni vele. Ez két, jelentéses egységnél nagyobb terminusok kiszűrésére már csak komplikáltabban használható. Ezen problémák megoldására szokták javasolni, hogy a többtagú terminusjelöltet több lehetséges helyen bontsuk szét, és ezekre az összetett párokra is nézzük meg, hogy mennyire koherensek. Ezek viszont bonyolult számításokat igényelnek, így olyanokat érdemes megvalósítani, amelyek nem két elemre épülnek.

A következő táblázatot Fahmi és mtsai (2007: 4) cikkéből vettük, amely nagyon jól összesíti a legtöbb, TE-ben is használt bigram metrikát:

Method	Formula
Frequency [15]	n_{11}
T-Score [5]	$\frac{n_{11}-\frac{n_{1p}n_{p1}}{n_{pp}}}{n_{11}^2}$
Log-likelihood [11, 7]	$2(n_{11}log\frac{n_{11}}{m_{11}}+n_{12}log\frac{n_{12}}{m_{12}}+ n_{21}log\frac{n_{21}}{m_{21}}+n_{22}log\frac{n_{22}}{m_{22}})$
Chi-squared (χ^2) [4]	$2(\frac{n_{11}-m_{11}}{m_{11}}^2+\frac{n_{12}-m_{12}}{m_{12}}^2+ \frac{n_{21}-m_{21}}{m_{21}}^2+\frac{n_{22}-m_{22}}{m_{22}}^2)$
Dice [10]	$2\frac{n_{11}}{n_{p1}+n_{1p}}$
Pointwise Mutual Information (PMI) [12, 5]	$log\frac{n_{11}}{m_{11}}$
True Mutual Information (TMI) [20]	$\frac{n_{11}}{n_{pp}}log\frac{n_{11}}{m_{11}}+\frac{n_{12}}{n_{pp}}log\frac{n_{12}}{m_{12}}+ \frac{n_{21}}{n_{pp}}log\frac{n_{21}}{m_{21}}+\frac{n_{22}}{n_{pp}}log\frac{n_{22}}{m_{22}}$
C-value [14]	$\begin{cases}\log_2\lvert a\rvert . f(a) & \text{if } a \text{ is not nested,}\\ \log_2\lvert a\rvert (f(a)- \frac{1}{P(T_a)}\sum_{b\varepsilon T_a} f(b)) & \text{otherwise}\end{cases}$

5.2. ábra. Bigram alapú statisztikai terminuskinyerő algoritmusok

5.4.3. Kontextus figyelése

Sok terminológiakivonatoló alkalmazás kidolgozója rájött arra is, hogy a terminusokat gyakran nem csak a belső szerkezetük alapján ismerhetjük fel, hanem az is nagy segítség lehet, hogy ha azt vizsgáljuk, milyen szavak állnak előttük, idézőjelben állnak-e stb. Ez azt jelenti, hogy a szövegkörnyezet is gyakran segíthet annak meghatározásában, hogy egy terminusjelölt tényleg az-e, vagy csak egy köznyelvi szó vagy szócsoport. Elég csak megvizsgálni az előtte és utána álló elemeket, vagy ez utóbbiak egy konkrét csoportját.

Frantzi és Ananiadou (1997) vizsgálták, hogy vannak olyan, főként igei szerkezetek, amelyek általában terminusokat vezetnek be. Például a *show* ige abban a szerkezetben, hogy *XY shows a basal cell carcinoma*, vagy a *known as* vagy a *called* abban a szerkezetben, hogy *which is called the Cartesian product*. Ha ilyen szerkezeteket találunk a szövegekben, akkor ezt nagy valószínűséggel terminus követi.

Hoste és mtsai (2008) is főleg erre épít: megnézi azokat a szavakat, amelyek általában terminusokat előznek meg vagy követnek (*is referred to, denotes, is called*). A vizsgálatot úgy végzik, hogy a program minden jelöltnél megnézi annak bal és jobb oldalán álló két szót/tokent (ez lehet zárójel is), és az alapján ad valószínűséget arra, hogy

az tényleg terminus-e. Sőt, a kontextus figyelése nemcsak a szavakra, hanem egyéb írásjelekre is kiterjed, mint a zárójelre is, hiszen zárójelben is állhat terminus, illetve egy zárójelben lévő kifejezés magyarázhat is terminust. Ők többféle információt és módszert használtak terminológiakivonatolásra, és azt tapasztalták, hogy a legjobb kombináció a lokális kontextus figyelembevétele a tokenek szófaji információival és lemmáival együtt. Tehát ezek azok az elemek, amelyekre mindenképpen szükség van, és ezt lehet még más egyéb statisztikai alkalmazásokkal javítani.

Yang és mtsai (2008) ezzel szemben azt állítják, hogy a legjobb, szinte egyetlen megoldás a terminológia kinyerésére az, hogy csak kontextuális információkat alkalmazunk ezen feladat megvalósításához. Ezt azzal indokolják, hogy a határolók (*delimiter*) használata egy stabil pont, amely doménfüggetlen. Nincs szükség a gyakoriság vizsgálatára, nincs szükség tanítókorpuszra sem, amelyet ráadásul minden főbb tudományterületre be kellene tanítani. Így könnyebben adaptálható bármilyen szakterületre, és olyan területeken is jól alkalmazható, ahol jelenleg nincs elég korpusz. Ezen kijelentések számunkra kissé szélsőségesek, de azt is tudnunk kell, hogy mindezt a kínai nyelvvel kapcsolatban állítják. A kínai nyelvben szerintük a POS-taggelés és a lemmatizálás amúgy is sokkal problematikusabb, így azon információk kevésbé használhatók. Erre egy egyszerű példát is hoznak, amelyet most csak angol fordításban közlünk: *Scan tunneling microscope is a kind of quantum-tunneling-effect based high angular resolution microscope* (Yang és mtsai 2008: 248). Ebben a mondatban a *Scan tunneling microscope, quantum-tunneling-effect* és a *high angular resolution microscope* a terminus. Ezek határolói az *a kind of*, a *based* és a kínai nyelvben létező melléknév-jelölő karakter. Itt már az is sokat elárul, hogy melléknév-jelölő karakter szinte csak a kínaiban van, így azzal nem tudunk mit kezdeni, mert a franciában ez nem létezik. Bizonyos mellékneveknél fellelhetők ugyan képzők a franciában (pl. *ration* 'értelem' → *rationnel* 'ésszerű', vagy *malheur* 'balszerencse' → *malheureux* 'szerencsétlen'), de nem minden melléknév képződik más szófajú szóból (pl. *beau* 'szép' vagy *franc* 'őszinte'). A kutatók körülbelül 80-90% körüli pontosságot értek el ezzel a módszerrel: konkrét szám azért nem adható meg, mert a pontosság annak függvényében változik, hogy hány delimitálót használtak: a túl kevés és a túl sok is kisebb pontossághoz vezet. Az algoritmusukat ők TCE_DI-nek, azaz *Term Candidate Extraction – Delimiter Identification*-nek nevezték el. A határolók listáját tanulókorpuszból nyerték ki, *stopword*-listák használatával.

Korábbi tanulmányainkban (pl. Nagy 2009a) megállapítottuk, hogy nagy előnnyel jár, ha ismerjük azokat a kifejezéseket, amelyek terminusok előtt szerepelnek. Azt láttuk, hogy az általunk akkor vizsgált korpuszban az *est appelé* (*is called* francia megfelelője), vagy a *notion de* 'valaminek a fogalma', vagy a *type de* 'valamilyen típusú' előtagok kifejezetten arra utalnak, hogy utánuk terminus helyezkedik el. De azt is észrevettük, hogy ezekkel igencsak vigyázni kell, mert ezek egy részhalmaza a terminusok része is lehet. Az akkori korpuszunk egy programozással kapcsolatos korpusz volt, így gyakran előfordultak olyan kifejezések, mint *type de donnée abstrait* 'absztrakt adattípus', amelyben a *type de* a terminus része, nem pedig egy olyan előtag, amely terminust vezet be. Ezért kell megkülönböztetni a határolók között is azokat, amelyek a terminus részei is lehetnek, mint ahogy az előbb említett *type de*, és azokat, amelyek sosem lehetnek terminus részei (pl. *est appelé*).

Nagy (2009a) azt is megemlíti, hogy nemcsak ezen elemek lehetnek hasznosak a TE esetén: különböző tipográfiai elemeket is célszerű figyelembe venni. Többek között azt, hogy az adott szó dőlt betűvel vagy félkövér karakterrel szerepel-e a szövegben, idézőjelben van-e, akár az angolszász típusú " " vagy a francia stílusú « » jelek között.

A legjobb tehát egy olyan módszer, amelyben figyelembe vesszük azt a lehetőséget is, hogy ezen előtagok a terminusok részei is, és attól függően, hogy milyen gyakorisággal szerepelhet egyik vagy másik terminusjelölő terminus közelében, attól függően rendelünk valószínűséget az utána következő elemhez. Ehhez elég jó támpontot ad Frantzi és Ananiadou (1997), valamint ennek továbbfejlesztett változata Maynard és Ananiadou (2000), ahol ezek a határolók automatikusan egy súlyértéket is tartalmaznak, amelyek azt mondják meg, hogy milyen valószínűséggel követi ezeket terminus. Ezeket a súlyértékeket az alábbi módon számolják ki: kiválasztják azon elemek egy részét a terminusjelölt-listában, amelyek biztosan terminusok, és ezeknek megnézik a környezetét, majd az alábbi képlettel kiszámolják annak súlyát, hogy az adott *w* kontextusszó (ami a biztos terminus környezetében áll) milyen valószínűséggel határolóegység:

A $$Weight(w) = 0{,}5 \cdot \left(\frac{t(w)}{n} + \frac{ft(w)}{f(w)} \right)$$ képletben a *w* a vizsgálandó kontextustoken (pl. határozott névelő, vessző stb.), *n* azon biztos terminusok egyszeri előfordulási száma, amelyekre ez a vizsgálat kiterjed (például ha 100 biztos, különböző terminusjelöltet választunk, akkor 100), *t(w)* azon esetek száma, ahol egy biztosnak jelölt terminus ezzel a *w* szóval áll együtt (az előző példánál maradva ez legfeljebb 100 lehet), *ft(w)* azt mutatja

meg, hogy a *w* szó összesen hányszor fordul elő terminusokkal együtt, *f(w)* pedig *w* korpuszbeli előfordulásainak száma.

Ezt követően Frantzi és Ananiadou (1997) ezt az értéket normalizálja, mert a fenti formula csak azt számolja ki, hogy egy adott kontextusszónak milyen a valószínűségi értéke. Azonban egy terminusjelöltet több szó is körbevehet, ezért szükség van egy olyan formulára is, amely megnézi egy adott szó teljes környezetét, és erre számol ki egy összes súly értékéket. Ennek a formulája a következő:

$$wei(a) = \sum_{b \in C_a} weight(b) + 1$$

Az *a* a vizsgált terminusjelölt, C_a *a* környezete, *b* egy környezetbeli szó, *weight(b)* pedig már az előbb is említett súly. Ezen érték kiszámolása után már csak a terminusjelölt tényleges *terminus technicus* valószínűségét kell kiszámítani, amelyben ez az érték 20%-ban számít bele, tehát ők úgy vélik, hogy ez az információ csak az egész valószínűségi érték egyötödét teszi.

Mindezek alapján a határolóelemek figyelembevétele nagyon fontos lehet, de tény, hogy teljes mértékben nem lehet erre támaszkodni Yang és mtsai (2008) állításával ellentétben, mivel ezen határolószavak listája szövegtípusonként is eltérhet. Az *est appelé, notion de* stb. kifejezések valószínűsíthetően gyakran fordulnak elő inkább vulgarizációs szövegváltozatokban, mint komolyabb szakszövegekben.

5.4.4. Tulajdonnevek szűrése

Grishman (2003) szerint a névelem-felismerés (a későbbiekben NER, azaz Named Entity Recognition) feladata a szövegben szereplő tulajdonnevek, pl. vállalatok, emberek, szervezetek nevei, földrajzi nevek, valamint megadott típusú entitások (például kémiai anyagok neveinek) felismerése. A szöveg típusától függően különböző típusú névelemek dominálnak: egy biológia szövegben például sok a fajmegnevezés, az újságokban pedig sok a vállalat-, hely- és személynév. Egy NER-alkalmazásnak nemcsak felismerni kell tudnia a névelemeket, hanem azokat szemantikai jellemzője alapján csoportosítani, tehát azt is meg kell adnia, hogy milyen típusú névelemről van szó. A névelem-felismerés szemléltetésére mutatunk egy példát, amely bemutatja egy NER-alkalmazás kimenetét egy adott szövegen:

> Nap Pál urat, közismertebb nevén Nap Palit, a Pásztorok Országos Érdekképviseletének elnökévé választották Komlószabolcson 6000 Ft kaució letételének kötelezettségével.

Az ehhez a mondathoz tartozó kimenet:

> <NAME TYPE=PERSON>Nap Pál</NAME> urat, közismertebb nevén <NAME TYPE=PERSON>Nap Palit</NAME>, a <NAME TYPE=ORGANISATION>Pásztorok Országos Érdekképviseletének</NAME> elnökévé választották <NAME TYPE= LOCATION>Komlószabolcson</NAME> 6000 <NAME TYPE=CURRENCY>Ft </NAME> kaució letételének kötelezettségével.

Egy NER-alkalmazást kétféleképpen lehet létrehozni: szabály alapú és gépi tanulási módszerekkel (Grishman 2003). A szabályok azt mondják meg, hogy a tulajdonnevek milyen mintára illeszkednek: például a nagybetűs írásmód alapkövetelmény a tulajdonnevek esetében, de nem elegendő. A magyarban egy szervezet nevének minden tagját nem mindig írjuk nagybetűvel, így legfeljebb az első tagról mondható el biztosan, hogy tulajdonnév. A németben pedig minden főnév nagybetűvel történő írása igencsak megnehezíti ezt a feladatot. A mondat eleji nagybetűs szó főleg nem mond el semmit annak tulajdonnévi mivoltáról. Megadhatunk olyan mintát is, amely például a személyneveket gyűjti újságokból: a nagybetűs szavak-vessző-kétjegyű szám-vessző karaktersorozat, amelyre például a *Nagy János, 42,* illeszkedik. Ezen kívül lehet még listát is létrehozni a lehetséges összes tulajdonnévről, de ez ezek száma miatt nem lehet hatékony megoldás. Ráadásul naponta keletkeznek új névelemek, így a frissítés is nehézkes.

A statisztikai módszerek esetén ez a probléma nincs meg: nem kell listát létrehozni lehetséges elemekről, csak egy nagyméretű tanítókorpuszra, és megfelelő algoritmusra van szükség. Ezen kívül, az ember is képes arra, hogy egy szövegben a névelemeket felismerje, anélkül, hogy azokat név szerint ismerné, így a névelemek felsorolása nem vezethet jó eredményre. Névelem-felismerést rengeteg tanuló módszerrel hajtottak már végre: van köztük olyan, amely rejtett Markov modellel (pl. Bikel és mtsai 1997), maximum entrópia modellel (Mikheev és mtsai 1998) vagy döntési fával (Sekine és mtsai 1998).

Mivel a névelemeknek rengeteg fajtája létezik, fontos rögzíteni, hogy milyen típusú névelemeket kell kiszűrni a TE során. Mivel például a kémiai névelemek nem tulajdonnevek, és ráadásul terminusok is lehetnek (pl. benzol), ezért azokat nem kell eltávolítanunk a terminusjelölt-listából. Elég csak a szövegben szereplő személy-, vállalat- és földrajzi nevek kiszűrése, amelyek például egy találmány leírásakor nem annak műszaki jellemzőit írják le, hanem a feltalálók vagy hivatkozott személyek neveit a vállalatuk megjelölésével. Mivel a leghatékonyabb megoldást a statisztikai módszerek jelentik (Grishman 2003), így a tulajdonnévszűrőt a statisztikai modulok fejezetbe tettük.

5.5. Terminológiakivonatolók összehasonlítása

Az előző két alfejezetben említett szabály alapú és statisztikai módszerek nem fedték le az összes, terminológiakivonatolásban is használt algoritmust, azonban megfelelő támpontnak bizonyulnak a saját terminológiakivonatoló létrehozásakor. Továbbá azt is figyelembe kell vennünk, hogy ezek között sok olyan van, amelyet nem lehet egyszerre választani, így arra kényszerülünk, hogy kiválasszuk, melyek azok, amelyekre nekünk szükségünk lesz. Ezekben nagy segítségünkre lehetnek korábbi tanulmányok, amelyek azt mérik fel, hogy ezek milyen hatékonysággal működnek.

Cabré és mtsai (2001), valamint Ha és mtsai (2008) szerint a hibrid rendszereknél az esetek többségében a terminusjelölt-lista felállításához statisztikai módszereket használnak, amelyet valamilyen nyelvi szűrővel szűrnek. Vizsgálódásaink alapján azonban sok esetben az első terminusjelölt-lista létrehozásában használják a szabály alapú megközelítést, és aztán ezt a listát szűrik különböző statisztikai módszerekkel (pl. Boulaknadel és mtsai 2008; Daille 1994). Erre vonatkozóan konkrét statisztikával nem rendelkezünk, de az az elfogadott nézet, hogy a szabály alapú megközelítések nagy fedéssel és kisebb pontossággal rendelkeznek, míg a statisztikai alapon működő alkalmazások nagyobb pontossággal, de kisebb fedéssel bírnak. A hibrid módszereknél viszont ez a megállapítás nem feltétlenül állja meg a helyét, mert a szabály alapú modelleken is sokat javíthatnak a statisztikai modulok, és az fordítva is igaz, tehát a statisztikai megközelítésen sokat javíthatnak a nyelvészeti szűrők.

Ami a statisztikai modulokat illeti, azokról már több olyan cikket is találhatunk, amelyek elég jól összemérik ezek hatékonyságát. Ezen kívül azt is figyelembe kell vennünk, hogy melyik az a módszer, amely minden területre és minden típusú terminológia kinyerésére alkalmas.

Általános nézet (pl. Zhang és mtsai 2008), hogy az egyszavas terminusokat külön kell venni a többszavas terminusoktól, és hogy legtöbbször ugyanazon statisztikai módszereket nem is alkalmazhatjuk a két különböző típusú terminusokra. Elég csak azt figyelembe venni, hogy az egyszavas terminusokra csak a *termhood* típusú mértékeket lehet használni, míg a többszavasok esetében a *unithood*-mértékeket is.

Fahmi és mtsai (2007) például bigramok vizsgálata során azt tapasztalták, hogy önmagában a gyakoriság vizsgálata vezet a legrosszabb eredményre. Náluk a C-érték közepes eredményt hozott, míg a *log-likelihood* ennél jobb, végül pedig a *khi-négyzet* próba bizonyult a leghatékonyabbnak.

Zhang és mtsai (2008) azon módszereket hasonlítják össze, amelyek nem tesznek különbséget az egy- és a többszavas terminusok között, azaz mindkettőt megpróbálják kinyerni, és amelyek nem dobnak el terminusjelölteket azért, mert azok csak ritkán fordulnak elő egy korpuszban. A vizsgálandó korpuszt két csoportra osztották: az egyik a biológiai és orvosi GENIA-korpusz, a másik pedig a wikipedia (tehát egy vulgarizált szöveg), amelyből kiválasztottak 1052, állatot leíró cikkszót. Azért választották a wikipediát is, mert szerintük szükség van olyan kivonatolókra is, amelyek nem csak biológiai vagy orvostudományi szakszövegekből nyernek ki terminusokat.

Azt a következtetést vonták le először, hogy a korpusz megválasztása igenis számít: a különböző algoritmusok nem működnek ugyanúgy a különböző korpuszokon. A *weirdness* például nagyon jól teljesített a wikipedia korpusz esetén (körülbelül 80%-os pontossággal), de a GENIA-korpuszban nagyon silány teljesítményt ért el: mindössze körülbelül 58%-ot. Ezzel szemben a C-érték és a tf-idf nagyon jó eredményt ért el a GENIA-korpuszon, körülbelül 80%-ot, de a wikipediás korpuszon körülbelül 60%-ot. Szerintük a C-érték azért is működhetett ilyen jól, mert azt főleg ilyen típusú szövegeken tesztelték, és mert főleg többszavas terminusokat kezel, és a GENIA-korpuszban a terminusok csak 11%-a egyszavas.

Boulaknadel és mtsai (2008) a különböző bigramokra alkalmazható statisztikai módszereket vette górcső alá, és mindezt az arab nyelvre alkalmazta többszavas terminusok kinyerésére. Az elemzett algoritmusok között szerepel a *log-likelihood ratio*, a t-érték és a Mutual Information is. Ők azt tapasztalták, hogy az adott környezetvédelmi korpuszukban a *log-likelihood* érte el a legjobb eredményt 85%-kal, a többi ehhez képest jócskán alulmaradt.

Hoste és mtsai (2008) arra jutott, hogy a TE esetében a kontextus figyelése a leghatékonyabb módszer. Azt tapasztalták még, hogy a tf-idf és a *log-likelihood* önmagukban az angolra ugyan nagy pontosságot hoznak (95%), de alacsony fedést (24%), azonban a hollandban ezek az értékek kiegyensúlyozottabbak lettek (65% pontosság, 63% fedés). Arra jutottak, hogy a tf-idf algoritmusa nagyon hasznosnak bizonyult, ha a terminusjelölt-lista szűréséről volt szó. A tf-idf<0,1 szűrő alkalmazásával a hatékonyság 8%-kal nőtt a többi tényező figyelembevételével kiszűrt terminusok közül. Mindez azt bizonyítja, hogy az ilyen statisztikai módszerek alkalmazása mindenképpen csak a terminusok szűrése esetében alkalmazható, a kinyerés esetében nem.

Lefever és mtsai (2009) elsősorban egy olyan alkalmazás kifejlesztését írja le, amelynek célja többnyelvű terminológia kinyerése, azaz a terminuskinyeréssel egyidőben egy többnyelvű szótár összeállítása is párhuzamosított korpuszból. Első körben három nyelvpárra tesztelték az alkalmazást: francia-angol, francia-olasz és francia-holland, és mindezt egy francia autóipari cég megbízásából, tehát a korpusz is ezt a területet tükrözte. A cikk tartalmaz egy részt, amely megmutatja, hogy az angol nyelvre milyen eredményeket ért el a program egynyelvű terminológiakivonatoló modulja. Itt az derült ki, hogy a C-érték 80%-os összeredményt ért el az egyszavas, míg 94%-ot a többszavas terminusok esetén. A *weirdness*-értéknél viszont azt állapították meg, hogy az egyszavas terminusok esetén sokkal jobban teljesített, mint társai, mégpedig 95%-kal.

5.6. Francia nyelvre készült terminológiakivonatolók

Az 5. fejezet eddigi részeiben a terminológiakivonatolók általános jellemzőit mutattuk be: a TE lépéseit, és a TE során használatos szabály alapú és statisztikai modulokat. Ezek ismertetése során a francia nyelvre kidolgozott módszereket is említettük, de nem hangsúlyoztuk ki, mert eddig a – szinte összes nyelvre alkalmazható – módszereket írtuk le. A jelen alfejezetben a legtöbbet idézett, francia nyelvre alkalmazható terminológiakivonatolókat mutatjuk be: ezek az Acabit (Daille 1994), a Lexter (Bourrigault 1994) és a Fastr (Jacquemin 2001).

Ezen kivonatolók közös jellemzője, hogy mind szabály alapú terminuskinyerést alkalmaznak. Így az a tény, hogy ezek a leginkább idézett TE-eszközök, alátámasztják azon hipotézisünket, miszerint a terminusok kinyerése a francia nyelv esetén szabály alapon is jól működhet. Ezen programok bemenetként már egy automatikusan annotált szöveget kapnak, amelyekben a mondathatárok, a tokenek, ez utóbbiak szótöve és szófaja már be van jelölve.

Az Acabit (Daille 1994) az előre annotált korpuszból főnév+melléknév, illetve főnév+főnév szekvenciákat nyer ki, de megengedi azt is, hogy köztük bármilyen prepozíció vagy determináns lehet. Ez nagyon előnyös lehet akkor, ha a terminusvariánsokat nem szeretnénk figyelembe venni: a *traitement*; *parole* 'kezelés, beszéd' páros például mind a terminust (*traitement de parole* 'beszédfelismerés'), mind annak variánsát (*traitement de la parole* 'a beszéd felismerése') egynek tekinti, és a későbbi statisztikai szűrő esetén a két külön variáns nem számít külön terminusnak. A főnevek és a melléknevek keresése annyiban előnyös még, hogy ezek a főnévi terminusok

fő jelentéses egységei. A szópárok kinyerése után a program statisztika segítségével szűri ezen szópárosokat: először csak a kettőnél nagyobb előfordulású szópárokat tartja meg, majd a megmaradt elemekre kiszámolja az 5.4.2.1. fejezeteben ismertetett MI (Mutual Information) értéket:

$$MI(x,y)=\log_2 \frac{P(x,y)}{P(x)P(y)}$$

A fenti képletben az x és y a vizsgált szópár két tagja. A kettő MI-értéke MI(x,y), P(x) az x szó előfordulásának valószínűsége, és P(x,y) az együttes előfordulás valószínűsége. A képletből az következik, hogy minél nagyobb az együttes előfordulás valószínűsége a két szó külön valószínűségének szorzatához képest, annál nagyobb MI-értéket kap, így annál nagyobb valószínűséggel lesz terminus.

A Lexter (Bourrigault 1994) a terminusok kinyeréséhez határolókat használ. Ez azt jelenti, hogy a terminusokban ilyen elemek nem találhatók, de megelőzhetik vagy követhetik, és ezáltal határolják azokat. Terminushatárolók többek között a ragozott ige, kötőszó, prepozíció+birtokos determináns. Például a *Les logiciels malveillants nuisent à nos ordinateurs* 'a rosszindulatú szoftverek kárt tesznek a számítógépeinkben' esetében a határolók (pl. prepozíció+birtokos determináns: *à nos)* kijelölése után megmarad a *logiciel malveillant* 'rosszindulatú szoftver' és az *ordinateur* 'számítógép' kifejezés, amelyek tényleg terminusok. A nem egyértelmű határolók esetén egy tanulókorpuszt használnak, amelyből kinyerhető az, hogy bizonyos szópárok (pl. *sur+le* 'a ...-n') milyen valószínűséggel részei terminusoknak (tehát mennyire produktívak), illetve milyen valószínűséggel határolók: a program a két valószínűség ismeretében tudja eldönteni, hogy az adott esetben ez határoló vagy sem.

A Fastr (Jacquemin 2001) alkalmazás célja elsősorban dokumentumok indexelése – azaz nem a TE – és az azokon belüli terminusvariánsok kezelése. A szövegben nemcsak terminusokat keres, hanem megállapítja azt is, hogy azok közül hány terminus szerepel a szövegben valamilyen más formában is. Ehhez az angol és a francia nyelven részletesen megfigyelt terminusvariációk alapján adtak meg szabályokat, amelyeket metaszabálynak neveztek el. Például az N_1 + de + N_2<$Lemma_2$> előfordulhat N_1 + A_1<$Lemma_2$> alakban is, ahol az azonos indexű elemek azonos karaktersorozatot jelölnek, így a terminusvariánsban a főnévi fejnek (N_1) meg kell egyeznie az eredeti terminus főnévi fejével, és a főnevet követő melléknév lemmájának meg kell egyeznie az eredeti terminus bővítményében szereplő főnév lemmájával. Ezen szabály alapján a *pression de sang* 'vérnyomás' terminusvariánsa a *pression sanguine* 'vérnyomás'. A terminusjelöltek és azok

variánsainak kinyeréséhez az annotált korpuszt használja fel a program. Ez alapján a bemenet alapján unifikációs nyelvtannal felismeri a terminusokat, és a nyelvtan segítségével egyben szintaktikailag is elemzi. Ezek után végzi el a terminusjelöltek variánsainak keresését, amelyekhez a metaszabályokat és egyéb tanult szabályokat alkalmazza.

5.7. Saját terminológiakivonatoló megvalósítása: célok és hipotézisek

Ebben az alfejezetben azt írjuk le, hogy a saját terminológiakivonatolónkat milyen elvek alapján hoztuk létre, mindezt az 5. fejezet alfejezeteire alapozva. Ebben a szakaszban még csak nagy vonalakban írjuk le az általunk felvázolt lépéseket, amelyeket aztán a következő fejezetekben bővebben kifejtünk.

A terminológiakivonatolónk célja, hogy az adott korpuszunkban meglevő terminusokat a lehető legpontosabban kinyerje: ez azt jelenti, hogy az adott korpuszból lehetőleg minden szakkifejezést a maga valódi hosszában kinyerjen. Ha egy adott terminusnak nem része az utána álló melléknév, akkor azt ne vegye a terminushoz a terminusjelölt listában, illetve ha többszavas, akkor az összes szó szerepeljen a terminuslista bejegyzésben, és ne külön-külön.

A korpuszt úgy választottuk ki, hogy francia nyelvű és erősen szakmai korpusz legyen, hogy tényleg a valódi terminusok szerepeljenek benne. A korpuszunk informatikai és élelmiszeripari szabadalmak leírását tartalmazza, és több különálló szövegből épül fel. Ez azonban nem jelenti azt, hogy a kivonatolónkat csak erre a területre szeretnénk megvalósítani: így nem szeretnénk olyan morfológiai információkat figyelembe venni, amely csak egy adott területre alkalmazhatók. Ilyen például a már korábban is említett biológiai szövegekben a *micro-, iso-* stb. előtagok, vagy informatikában a francia nyelvű *-ciel* végű főnevek, mint *graticiel* (’ingyenes program/freeware’), vagy *partagiciel* (’demóverzió’).

A bemeneti korpusz mondatokra, tokenekre bontását, illetve azok szótövesítését és szófaji címkézését nem saját modul hajtotta végre, hanem az erre a célra szolgáló, Connexor vállalat Machinese szintaktikai elemző modulja. Ezen program az összes olyan információval ellátja a szöveget, amelyre szükségünk volt, így további nyelvi elemzést végrehajtó modult nem építettünk be.

Mivel főbb célunk a terminusok kinyerésekor a minél nagyobb fedés elérése, még ha ez kissé a pontosságon ront is, ezért a terminusok legelső kinyerésére szabály alapú

módszereket alkalmaztunk. Ez azt jelenti, hogy kézzel megadott szabályok alapján a már felcímkézett szövegekből az alkalmazás kinyerte az elsődleges terminusjelölt-listát. Mivel a véges állapotú automatával történő kinyerés számunkra átláthatóbb, mint a reguláris kifejezéseké, így ezt a módszert választottuk. Miután illesztettük a mintát, kinyertük az összes olyan lehetséges karaktersorozatot, amely esetleg terminus lehet.

Mivel a szakirodalom szerint a szabály alapú módszerek nagy fedést érnek el alacsony pontosság mellett, így az az előfeltételezésünk, hogy a szabály alapú terminuskinyeréssel nagy fedést érünk el alacsony pontossággal.

A szabály alapú kinyerés hatékonyságának növelése érdekében szabály alapú szűrést alkalmaztunk a szabály alapú kinyerés előtt. Ezt egy *stopword*-listával hajtottuk végre, amely tartalmazott tulajdonneveket és olyan főneveket, mellékneveket, amelyek nem lehetnek terminusok részei (ezek a konnektívumok). Ezáltal az volt az előfeltételezésünk, hogy a pontosság jelentősen növekedni fog, mert így sok olyan terminusjelöltet is kiszűrünk, amely valószínűleg nem terminus.

Ezt a listát a későbbiekben statisztikai módszerekkel tovább szűrtük, amihez szükségünk volt egy *termhood*- valamint egy *unithood*-algoritmusra. A többszavas terminusok esetén egy *unithood*-algoritmust kellett végrehajtani, amellyel meghatároztuk, hogy a több egységből álló szakkifejezések mely szegmenseit kell esetleg elvetnünk, tehát azt kellett meghatároznunk, hogy a karaktersorozat mely részét (amely akár az egész is lehet) tartottuk meg terminusnak. Ehhez a már korábban is említett C/NC-érték algoritmust (Frantzi és Ananiadou 1997) választottuk, amely egyben a kontextusokat is elemzi, és így azt is meg tudja mondani, hogy a szövegkörnyezet alapján milyen valószínű, hogy az adott karaktersorozat terminus. A C-érték választását az indokolta, hogy Zhang és mtsai (2008) szerint egy szaknyelvi korpuszon, a GENIA-korpuszon, az bizonyult a leghatékonyabbnak.

A terminusok kontextusait figyelembe vevő értékének ezért a C-értékhez kapcsolódó, a C/NC-érték kiszámítására szolgáló súlyértéket (*Weight*) használtuk (Frantzi és Ananiadou 1997). A súlyérték a terminusjelölt környezetében előforduló lehetséges tokeneknek ad valószínűségi értéket, majd egy adott terminusjelöltnek egy másik valószínűségi értéket az adott környezetében lévő tokenek értékei alapján. Annak ellenére, hogy a *Weight*-érték nem tesz különbséget a terminusjelölt előtti és utáni tokenek között (például a terminusjelölt előtt lévő vessző ugyanolyan súlyértékkel rendelkezik, mintha utána állna), a saját terminológiakivonatolóban ezt használtuk, mert először a C/NC összevont értéket szerettük volna használni (amely erre a súlyértékre épül), de mivel egy

harmadik statisztikai mértéket is használtunk, ezért később csak a C/NC komponenseit alkalmaztuk.

Mivel nem rendelkeztünk referenciakorpusszal, azaz egy olyan szövegtárral, amely általános nyelvi szövegeket tartalmaz, így olyan *termhood*-értéket kellett választanunk, amelyhez ez nem szükséges. Ezen típusú értékeket nemcsak a többszavas, hanem az egyszavas terminusokra is ki kellett számolnunk: meg kellett állapítani azt az arányt, ahogy a köznyelvben és a szaknyelvi korpuszban előfordulnak. Mint ahogy a 7.5.1. fejezetben kifejtjük, adott terminusjelölt köznyelvi korpuszbéli előfordulási arányának kiszámításához egy internetes keresőmotort használtunk, amelyből csak azt tudjuk meg, hogy az adott terminusjelölt hány dokumentumban fordul elő, de azt nem, hogy az egyes dokumentumokban hányszor. Ezért olyan statisztikai mértéket kellett használnunk, amely adott terminusjelölt dokumentumok számára vonatkoztatott előfordulási számát adja meg (azaz összesen hány dokumentumban fordul elő, de lényegtelen, hogy abban hányszor). Ezért az 5.4.1.3. fejezetben kifejtett *weirdness*-értéket alkalmaztuk. Mivel ez az érték, amely az adott terminusjelölt szaknyelvben és a köznyelvben lévő előfordulási arányának a hányadosa, így egyszerűen megragadja a *termhood*-értékek lényegét: azaz ha egy kifejezés egy adott szakterületen nagyobb arányban fordul elő, mint egy általános nyelvi korpuszban, akkor az nagyobb valószínűséggel lesz terminus, és ezen mérték alapján nagyobb *weirdness*-értéket fog kapni.

Mivel a megfelelően megválasztott statisztikai módszerek nagy pontosságot érnek el alacsony fedés mellett, és a szabály alapú kinyerés már eleve nagy fedést érhet el, így az előfeltételezésünk az, hogy a fenti statisztikai módszerek használatával a fedés lehető legkisebb csökkenése mellett a pontosság tovább növelhető azon jelöltek szűrésével, amelyek például gyakran fordulnak elő köznyelvben, vagy amelyeknek elemei nem tartoznak annyira egybe.

6. A korpusz bemutatása

Korpuszként olyan francia nyelvű szabadalmakat vettünk alapul, amelyek online elérhetők a WIPO (*World Intellectual Property Organization* – Szellemi Tulajdon Világszervezete) honlapjáról. A WIPO az ENSZ specializált ügynöksége, feladata egy olyan rendszer kidolgozása, amely a szellemi tulajdonnal foglalkozik, és mindenki számára elérhető (WIPO 2004).

6.1. A szabadalmak részei

6.1.1. Bibliográfiai adatok

A bibliográfiai adatok a szabadalmakhoz kapcsolódó legfontosabb dátumokat, számokat, neveket (pl. feltaláló neve) tartalmazzák, ez a rész a szabadalmak legelső része. A legfelső sorban található az adott szabadalom kódja, ami a különböző szabadalmi rendszerek miatt eltérő formátumú is lehet. A bemutatott példában nemcsak a WIPO-féle kód jelenik meg, hanem a nemzetközi szabadalmi kódszám is. Ebben a felső fejlécben olvasható a nyilvántartásba vétel dátuma, valamint az adott szabadalom publikálásának dátuma.

Az IPC-mezőben azok a kódszámok találhatók, amelyek segítségével azonosítható, hogy milyen területet érint az adott benyújtott szabadalom. Ez a kód minden esetben egy betűvel kezdődik, amely a főbb kategóriát határolja be. A betűk jelentése a következő:

A — Emberi szükségletek
B — Végrehajtó műveletek; szállítás
C — Kémia; kohászat
D — Textil; papír
E — Rögzített szerkezetek
F — Mechanika; világítás; fűtés; fegyverek; robbanóanyag
G — Fizika
H — Elektromosság[23]

Ezen belül a többi szám/betű az adott terület egy alfaját írja le. A példában ez a szabadalom az alábbi területhez köthető:

[23] A — HUMAN NECESSITIES
B — PERFORMING OPERATIONS; TRANSPORTING
C — CHEMISTRY; METALLURGY
D — TEXTILES; PAPER
E — FIXED CONSTRUCTIONS
F — MECHANICAL ENGINEERING; LIGHTING; HEATING; WEAPONS; BLASTING
G — PHYSICS
H — ELECTRICITY

G06F: digitális adat feldolgozása
G07C: […] gépek működésének regisztrálása vagy jelzése; véletlenszám-generálás; ellenőrző rendszerek (személyazonosítás, pl. ujjlenyomat, láblenyomat stb.)
H04L: digitális adat továbbítása

Ha megnézzük a szabadalom címét (*Biometrikus adatok feldolgozása transzformációval*), mindez világossá válik: a biometrikus adatok személyazonosításra szolgálnak (G07C), digitális adatok feldolgozásával (G06F) és azok továbbításával (H04L) járnak. Az IPC-kód alatt található a szabadalmat benyújtók, illetve a következő sorban a feltalálók neve (a kettő között általában van átfedés).

Pub. No.:	WO/2009/004215	International Application No.:	PCT/FR2008/051044
Publication Date:	08.01.2009	International Filing Date:	12.06.2008

IPC:	***H04L 9/32*** (2006.01), ***G06F 21/00*** (2006.01), ***G07C 9/00*** (2006.01)
Applicants:	**SAGEM SECURITE** [FR/FR]; Le Ponant de Paris, 27 rue Leblanc, F-75015 Paris (FR) *(All Except US)*. **CHABANNE, Hervé** [FR/FR]; (FR) *(US Only)*. **BRINGER, Julien** [FR/FR]; (FR) *(US Only)*.
Inventors:	**CHABANNE, Hervé**; (FR). **BRINGER, Julien**; (FR).
Agent:	**Cabinet Plasseraud et al.**; 52 rue de la Victoire, F-75440 Paris Cedex 09 (FR) .
Priority Data:	0704170 12.06.2007 FR
Title:	(EN) PROCESSING OF BIOMETRIC DATA BY TRANSFORMATION (FR) TRAITEMENT DE DONNEES BIOMETRIQUES PAR TRANSFORMATION

Ezt követően olvashatjuk a szabadalom angol és francia nyelvű absztraktját. Ebben a részben ábrákat is el lehet helyezni. A legvégén az iktatás, valamint a publikáció nyelve található.

Abstract:

(EN) Biometric data relating to a biological part are processed by obtaining, on the one hand, a first set of transformed biometric data (f(B1)) by applying at least one irreversible transformation to a first set of biometric data (B1), and, on the other hand, a second set of transformed biometric data (f(B2)) by applying said transformation to a second set of biometric data (B2).[…]

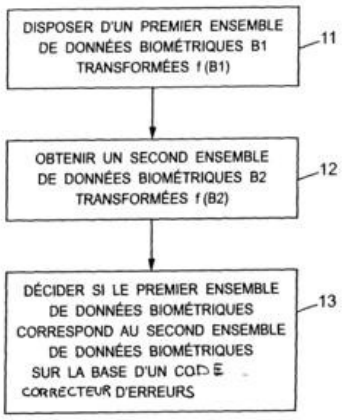


FIG. 1

11 ACQUIRE A FIRST TRANSFORMED f (B1) BIOMETRIC DATA SET (B1)
12 OBTAIN A SECOND TRANSFORMED f (B2) BIOMETRIC DATA SET (B2)
13 DECIDE WHETHER THE FIRST BIOMETRIC DATA SET CORRESPONDS TO THE SECOND BIOMETRIC DATA SET ON THE BASIS OF AN ERROR CORRECTOR CODE

(FR) Des données biométriques relatives à une partie biologique sont traitées en obtenant, d'une part, un premier ensemble de données biométriques transformées (f(B1)) par application d'au moins une transformation irréversible à un premier ensemble de données biométriques (B1), et, d'autre part, un second ensemble de données biométriques transformées (f(B2)) par application de ladite transformation à un second ensemble de données biométriques (B2). […]

Designated States: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, […].

Publication Language: French (FR)

Filing Language: French (FR)

6.1.2. Leírás és igénypontok

A leírás a szabadalom legterjengősebb része, célja a találmány pontos és részletes leírása. Ennek következtében igen hosszú, és rengeteg szóismétlés jellemzi annak érdekében, hogy még jobban körülírja az adott szabadalmat. A leírás a találmánynak mind az általános jellemzőit, mind annak specifikusságait tartalmazza. Mivel a leírásnak minél akkurátusabbnak kell lennie, ezért az összes olyan terminus megtalálható benne, amely az adott szabadalom területéhez köthető.

A leírást követő igénypontok tekintetében sokkal több megszorítás létezik. Legfontosabb, hogy legalább egynek kell lennie, ezt nevezzük főigénypontnak. Az igénypontnak pontosan rá kell mutatnia arra, hogy mit szeretne az igénylő szabadalmaztatni, és mindezt explicit módon, egy mondatban le kell írnia. Az igénypontok azok, amik jogi esetekben mérvadók, ezért ezeket ennek megfelelően kell megfogalmazni.

Az igénypont csakis egymondatos lehet, és az első részben tartalmaznia kell azt, hogy az illető pontosan mit szeretne levédetni (pl. módszer vagy eszköz), majd azokat kell bőven kifejtenie (Osenga 2006).

Az előző példánk főigénypontja a következő:

> 1. Procédé de traitement de données biométriques relatives à une partie biologique, ledit procédé comprenant les étapes suivantes :
> /a/ obtenir un premier ensemble de données biométriques transformées (f(B1)) par application d'au moins une transformation irréversible à un premier ensemble de données biométriques (B1) ;
> /b/ obtenir un second ensemble de données biométriques transformées (f(B2)) par application de ladite transformation à un second ensemble de données biométriques (B2) ;
> Ici décider si le second ensemble de données biométriques correspond au premier ensemble de données biométriques sur la base d'une comparaison entre le premier ensemble de données biométriques transformées et le second ensemble de données biométriques transformées ; ' dans lequel ladite comparaison est effectuée sur la base d'une représentation numérique desdits premier et second ensembles de données biométriques transformées en fonction d'un mot de code correcteur d'erreurs.

Az igénypont első része írja le, hogy egy eljárást védetne le (biometrikus adatok feldolgozása), amely az /a/ és /b/ pontban leírt lépéseket tartalmazza. A felsorolás után írja le, hogyan történik a transzformált és az eredeti biometrikus adatok összehasonlítása.

6.2. A kiválasztott korpusz

A szabadalmak részei közül a legrészletesebbet és a leghosszabbat választottuk korpuszként, azaz a leírást. A leírási rész hosszúsága és kiemelt pontossága azt vonja maga után, hogy a terminusok gyakran ismétlődnek benne, és legtöbbször ugyanabban a formában, így – feltehetően – hatékonyan alkalmazhatóak rájuk a statisztikai módszerek is. Az általunk választott leírások körülbelül 4000 tokent, azaz szövegszót tartalmaznak: ebbe nem tartoznak bele a html-tagek és az írásjelek. A html-tagek a szöveg formázását írják csak le, például az és tagek közötti rész dőlt betűsként jelenik meg. Az írásjelek közül a pont, a vessző stb. az általunk választott tokenizáló program szerint külön tokenek, de a fenti tokenszámba ezeket nem vettük bele.

A szabadalmi osztályok közül a G06F IPC-osztályt választottuk, ami a digitális adat feldolgozásának területébe tartozó szabadalmakat tartalmazza: a korpuszt alkotó legtöbb szabadalmi leírás ebből az osztályból származik. Ezt követően kiválasztottunk egy másik osztályt is azért, hogy megnézzük, hogy a terület kiválasztása milyen összefüggésben áll a terminológiakivonatoló hatékonyságával. Az általunk választott másik osztály az A23L osztály lett, amelyből az A jelentése emberi szükségletek, A23 az élelmiszereket és azok kezelését jelenti, az A23L ezen belül élelmiszerekkel, élelmiszerfélékkel vagy azon

alkoholt nem tartalmazó italokkal foglalkozik, amelyeket az A21D vagy az A23B-A23J alosztályok nem fednek le. Ez az osztály foglalkozik még az előbbiek elkészítésével vagy kezelésével, például a főzés, tápanyagminőség módosítása, fizikai kezelés, tartósítás.

Ez azt jelenti, hogy fő korpuszként informatikai leírásokat választottunk, de azt is szeretnénk megnézni, hogy például az élelmiszeriparhoz kapcsolódó szövegek esetében is alkalmazható-e az elsősorban informatikai dokumentumokra kifejlesztett terminuskinyerő alkalmazás. Ezzel azt is megtudjuk, van-e esetleg olyan minta, amely más típusú szövegekben előfordulhat, azaz mennyire szövegfüggő a mintaillesztésen alapuló terminuskinyerés.

Fontos megjegyezni, hogy a terminológiakivonatoló alkalmazás bármennyi szabadalomra lefut, de a korpusz méretét meg kellett határozni, mert a program validálási része csak akkor fut le egy adott szabadalmi leírásra, ha az abban fellelhető összes terminus kézzel be van jelölve. Erre azért van szükség, mert csak így állapítható meg, hogy az adott leírásban mennyi az alkalmazás fedés, pontosság és F-értéke.

Mindkét szabadalmi területről 10 leírást választottunk korpusznak, így összesen 20 leírás állt rendelkezésünkre. A választott szabadalmak kódját és címét (mind az eredeti francia nyelvű valamint a magyarra fordított címet) a 6.1. és 6.2. táblázat tartalmazza. A címek fordításai saját fordítások.

6.1. táblázat: A korpusznak kiválasztott 10 leírás adatai a G06F szabadalmi területről

Sor-szám	Dokumentum kódja	Tokenek száma	Francia cím	Magyar fordítás
1.	FR2008051044	3796	Traitement de données biométriques par transformation	Biometrikus adatok feldolgozása transzformációval
2.	FR2008051104	3665	Procédés et dispositifs pour la communication de données de diagnostic dans un réseau de communication temps réel	Diagnosztikai adatok valós idejű kommunikációs hálózatban történő közlésére létrehozott módszerek és eszközök
3.	FR2008051812	4491	Dispositif d'affichage d'une pluralité de documents multimédia	Számos multimédia-dokumentum megjelenítésére szolgáló eszköz
4.	FR2008051823	2760	Méthode et système d'annotation de documents multimédia	Multimédia-dokumentumokat annotáló eszköz és rendszer
5.	FR2008051836	3445	Échange de données entre un terminal de paiement électronique et un outil de maintenance par une liaison USB	Egy elektronikus fizető terminál és egy karbantartási eszköz USB-kapcsolaton keresztüli adatcseréje
6.	FR2008051856	3002	Système et procédé d'échange d'informations dans un terminal multimédia	Multimédia-terminálokban használható információcsere rendszer és eszköz
7.	FR2008051890	1767	Dispositif électrique à télécommande sans fil et à consommation réduite	Vezeték nélküli, távolról irányítható és csökkentett fogyasztású elektronikus eszköz
8.	FR2008052025	3276	Procédé et dispositif pour commander l'affichage d'une zone d'informations sur l'écran d'accueil d'un terminal mobile	Egy mobilterminál főképernyőjén lévő információs zóna megjelenítését vezérlő módszer és eszköz
9.	FR2008052073	3583	Vérification de données lues en mémoire	Memóriába beolvasott adatok ellenőrzése
10.	FR2008052077	5340	Système d'interprétation simultanée automatique	Automatikus szinkrontolmács-rendszer

6.2. táblázat: A korpusznak kiválasztott 10 leírás adatai az A23L szabadalmi területről

Sor-szám	Dokumentum kódja	Tokenek száma	Francia cím	Magyar fordítás
1.	EP2008056887	2726	Procédé de traitement de crustacés, par exemple de langoustines, en vue de leur conservation	Rákfélék, például languszták, eltarthatóságát biztosító eljárás
2.	EP2008061497	4286	Procédé de fabrication d'un système de délivrance de principes actifs d'origine végétale	Növényi eredetű aktív alapanyagok kinyerésére szolgáló rendszer gyártásának eljárása
3.	EP2008063597	3698	Procédé de compaction d'une composition pulvérulente à volume constant	Egy állandó térfogatú, por állagú összetétel tömörítési eljárása
4.	EP2009057808	5629	Procédé et dispositifs de fabrication de portions de produits alimentaires fourrées d'une garniture	Töltött élelmiszeradagok gyártási eljárása és eszközei
5.	FR2006001856	1130	Confiserie de chocolat d'aspect métallique	Ezüst hatású csokoládé alapú édességek
6.	FR2007001526	2513	Procédé et dispositif de préparation de produits naturels cristallisés par enrobage de sucre	Cukorbevonattal kristályosított természetes termékek készítésének eljárása és eszköze
7.	FR2007051158	4372	Utilisation du safran et/ou du safranal et/ou de la crocine et/ou de la picrocrocine et/ou de leurs dérivés en tant qu'agent de satiété pour le traitement de la surcharge pondérale	Sáfrány és/vagy safranal és/vagy krocin és/vagy pikrokrocin és/vagy származékaik használata jóllakottság érzését keltő szerként a túlsúly kezelésére
8.	FR2007051178	1515	Utilisation d'un mélange de neige carbonique et d'azote liquide dans des applications de surgélation	Szilárd szénsav és folyékony nitrogén elegyének használata fagyasztásra szolgáló eszközökben
9.	FR2007051372	8390	Ligne et procédé de traitement thermique de produits contenus dans des poches, avec scellement de ces dernières	Fóliában tárolt termékek hőkezelési eljárása az előbbiek lezárásával
10.	FR2007051549	3243	Système d'injection de fluide cryogénique permettant le traitement de produits en vrac et procédé de refroidissement le mettant en œuvre	Ömlesztett áruk kezelését biztosító fagyasztó folyadék befecskendezési rendszer, és az azt megvalósító hűtési eljárás

6.3. A korpusz kinyerése

Szerencsére a szabadalmak az interneten ingyenesen hozzáférhetők: ennek legegyszerűbb oka az, hogy bárki utána tudjon nézni, hogy létezik-e már szabadalom a keresett témában. Másik, még kézenfekvőbb ok az, hogy mielőtt levédetnénk valamit, elő tudjuk keresni, hogy azt már szabadalmaztatták-e vagy sem, így felesleges jogi költségektől tudjuk magunkat megvédeni.

Az interneten több forrásból is kereshetünk szabadalmakat. A Google keresőmotorban már nemcsak honlapokra kereshetünk rá, hanem egyéb típusú dokumentumokot is lekérhetünk, ugyanis ma már külön kereshetünk videókra, képekre, hírekre stb. Mindezek mellett a Google elindította még a szabadalomkeresési szolgáltatását is, amely elérhető a http://www.google.com/patents oldalról. Ez az oldal azonban jelenleg csak az Amerikai Egyesült Államokban iktatásba vett szabadalmakat tartalmazza, azokat is csak PDF formátumban, így azok számunkra nem használhatók. Szintén amerikai szabadalmakra lehet keresni a http://patft.uspto.gov/ oldalon, ahol két külön keresőablak áll rendelkezésre a benyújtott és elfogadott szabadalmak keresésére. Itt a szabadalmak már szöveges formátumban is elérhetők, de csak angol nyelvű szabadalmakra kereshetünk.

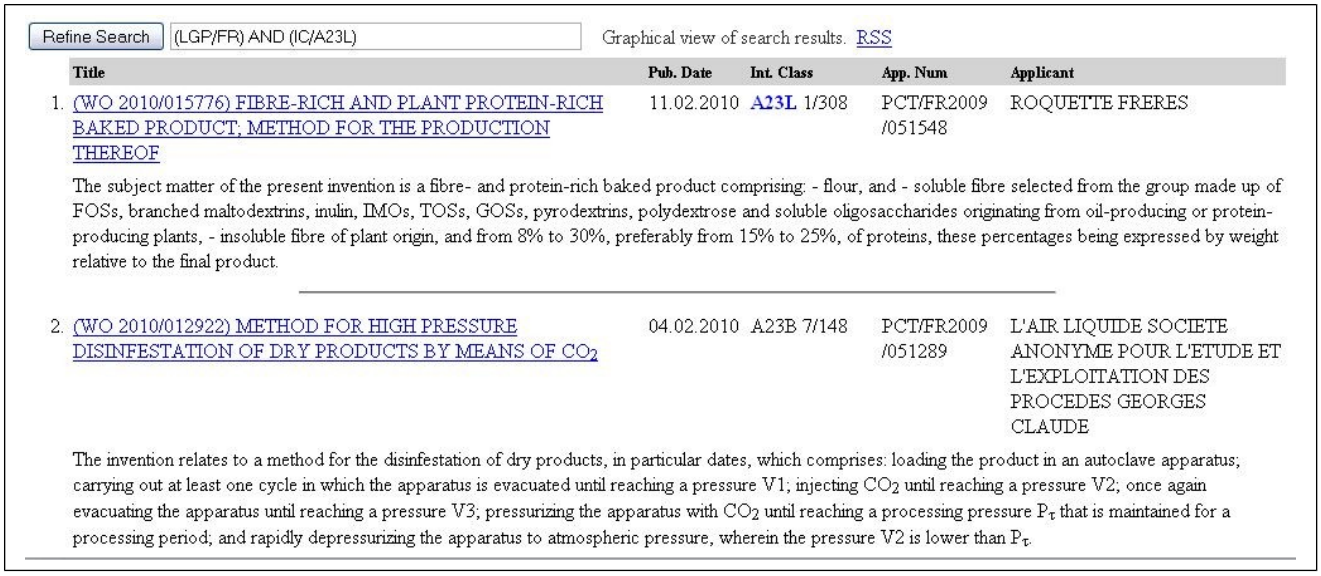
A vizsgálatban a WIPO szervezete által biztosított szabadalmi keresőt, a Patentscope-t használtuk (http://www.wipo.int/pctdb/en/), ahol nemzetközi adatbázisban szinte bármilyen nyelven vagy paraméter alapján kereshetünk. A választható paraméterek – többek között – az alábbiak lehetnek:

A szabadalom nyelve
IPC-osztály
Iktatás ideje
Cím
Feltaláló neve
Szabadalmaztató neve
Keresés igénypontokban
Keresés leírásokban

A keresés paramétere mellett még beállíthatjuk a megjelenítendő találatok számát, sorrendjét, és hogy a szabadalmak mely részeit jelenítse meg a találati oldal minden egyes találatnál.

A szabadalmi korpusz összeállításakor a keresőbe annyit írtunk be, hogy a publikálás nyelve francia legyen (a *Language of publication* értékét FR-re állítottuk), valamint azt, hogy informatikai, valamint élelmiszeripari szabadalmakra keressen (az Int.

Class értékét az első kereséskor G06F-re, másodjára A23L-re állítottuk), a megjelenítendő találatok számát pedig a maximálisra, azaz 500-ra állítottuk. A találati oldal első két rekordját a 6.1 ábra mutatja. Minden egyes találatnál megtalálható annak címe, megjelenési dátuma, a nemzetközi osztály és az absztrakt szövege.


Refine Search (LGP/FR) AND (IC/A23L) Graphical view of search results. RSS

	Title	Pub. Date	Int. Class	App. Num	Applicant
1.	(WO 2010/015776) FIBRE-RICH AND PLANT PROTEIN-RICH BAKED PRODUCT, METHOD FOR THE PRODUCTION THEREOF	11.02.2010	A23L 1/308	PCT/FR2009 /051548	ROQUETTE FRERES
	The subject matter of the present invention is a fibre- and protein-rich baked product comprising: - flour, and - soluble fibre selected from the group made up of FOSs, branched maltodextrins, inulin, IMOs, TOSs, GOSs, pyrodextrins, polydextrose and soluble oligosaccharides originating from oil-producing or protein-producing plants, - insoluble fibre of plant origin, and from 8% to 30%, preferably from 15% to 25%, of proteins, these percentages being expressed by weight relative to the final product.				
2.	(WO 2010/012922) METHOD FOR HIGH PRESSURE DISINFESTATION OF DRY PRODUCTS BY MEANS OF CO_2	04.02.2010	A23B 7/148	PCT/FR2009 /051289	L'AIR LIQUIDE SOCIETE ANONYME POUR L'ETUDE ET L'EXPLOITATION DES PROCEDES GEORGES CLAUDE
	The invention relates to a method for the disinfestation of dry products, in particular dates, which comprises: loading the product in an autoclave apparatus; carrying out at least one cycle in which the apparatus is evacuated until reaching a pressure V1; injecting CO_2 until reaching a pressure V2; once again evacuating the apparatus until reaching a pressure V3; pressurizing the apparatus with CO_2 until reaching a processing pressure P_T that is maintained for a processing period; and rapidly depressurizing the apparatus to atmospheric pressure, wherein the pressure V2 is lower than P_T.				

6.1. ábra. Találat oldal első két rekordja

Miután ezt a két oldalt elmentettük, minden automatikusan történik. A két HTML-oldalt a program beolvassa, és feladata az, hogy az ábrában kékkel jelölt hivatkozásokhoz tartozó URL-címet kinyerje, majd az ahhoz a címhez tartozó dokumentumot (szabadalmi leírást) szöveges fájlként elmentse.

Mivel ezek az oldalak HTML-formátumúak, ezért szükséges ezek egyszerű szöveges fájlokká történő konvertálása. Ezenkívül fontos, hogy ezen fájlokban csak az egyszerű szövegeket tartsuk meg, az oldal szélén található hivatkozásokat, felesleges fejléceket töröljük. A HTML-oldalak egyszerű szöveggé történő konvertálása azért szükséges, mert azokban például az *é* karakter *é* formában szerepel, és a későbbiekben a könnyebb olvashatóság kedvéért nem ezt választjuk, hanem a hagyományos írásmódot.

A HTML-oldalak felesleges adatainak eltávolítása már az adott oldaltól függ. A leírást tartalmazó oldalak esetében például a leírás a <description> és </description> tagek között szerepelnek, az igénypontok pedig a <claims> és </claims> tagek között, így ezeket tartottuk csak meg. Ezen tagek közötti részből még el kellett távolítani azokat a tageket is, amelyek a szöveg formázását valósítják meg.

6.4. A korpusz annotálása

A terminuskinyerési folyamat végén szükséges az alkalmazott algoritmus(ok) hatékonyságának mérése. Mivel a terminuskinyerési folyamat elég összetett (tokenizálás, kinyerés, szűrés stb.), az is előfordulhat, hogy az első verziókban a jobb eredmények érdekében egy-egy alfolyamatba bele kell nyúlni, és minden egyes javítás végén kíváncsiak vagyunk, hogy egy-egy változtatás mennyire módosítja az alkalmazás hatékonyságát a helyesen kinyert terminusok és a zaj tekintetében. Mivel minden egyes javítás után az alkalmazás által összeállított lista kézi áttekintése nem hatékony, mert például többször el lehet nézni vagy számolni, egy olyan kézzel bejelölt listára van szükség, amely tartalmazza a vizsgált dokumentumban szereplő terminusokat, így a számítógép automatikusan le tudja ellenőrizni, hogy a program mekkora fedést, pontosságot ért el az adott verziójában. Mivel a letöltött dokumentumokban a terminusok nincsenek jelölve, így a terminusok listáját kézzel hoztuk létre.

A terminusok kézi bejelölésekor a szöveges, nem szótövesített változatot vettük alapul, amelyekben csak a terminusokat hagytuk benne, a többi részt töröltük. Minden egyes terminus új sorba került, tehát egy szöveges fájlban tároltuk azok listáját. Fontos megjegyezni, hogy ezekben a fájlokban kézi szótövesítést hajtottuk végre, azaz minden egyes terminus minden egyes szava lemmatizált alakban került fel. Ez nem egyezik meg minden esetben a szótári alakkal, ahol általában csak a fej van szótövesítve. Erre példa a *base de données* 'adatbázis' vagy a *système d'exploitation* 'operációs rendszer', amelyek a terminusok listájába a nem létező *base de donnée* vagy *système de exploitation* alakban kerültek fel.

Azt, hogy mi terminus és mi nem, a 3. fejezet alapján döntöttük el: megfelel-e a klasszikus wüsteri definíciónak vagy sem. Ezen kívül segítségünkre volt a *Grand dictionnaire terminologique* online terminológiai szótár[24], ami francia-angol-latin viszonylatban tartalmaz terminusokat, szinte az összes tudományterületről, köztük azon doménekről is, mint az általunk kiválasztott informatika vagy élelmiszeripar.

[24] http://www.oqlf.gouv.qc.ca/ressources/gdt.html

7. Terminológiakivonatolás megvalósításának módszere

E fejezet célja annak bemutatása, hogyan jutunk el egy nyers szövegből terminusainak kinyeréséig. Ehhez az 5.2. alfejezetben leírtaknak megfelelően hajtjuk végre a TE lépéseit. Először a nyers szöveget egy programmal annotáltatni kell: mondatokra kell bontani, majd azokat tokenekre, valamint minden tokenhez hozzá kell rendelni annak szótövét és nyelvtani kategóriáját (7.1.). A továbbiakban csak ezeket a tokeneket vettük figyelembe; először szűrőket alkalmaztunk a tulajdonnevek és a főnévi fejet is tartalmazó szerkezetekre, hogy azok semmiképpen se kerüljenek a terminusjelölt-listába (7.2.). Ezt követően a ki nem szűrt tokenekre véges állapotú automatát illesztettünk (7.3. és 7.4.), majd az így kinyert terminusjelölteket statisztikai módszerekkel tovább szűrtük (7.5.).

7.1. A korpusz előfeldolgozása

A terminuskinyerési folyamat során szükségünk van az 5. fejezetben említett előfeldolgozási lépésekre. A bemeneti szöveget először mondatokra, majd tokenekre kell bontani, majd minden egyes tokenhez egy szófaji címkét, illetve szótövet kell rendelni. Ezt egy külső programmal vittük véghez.

7.1.1. Korpusz mondatokra, majd tokenekre bontása, és azok címkézése

A korpusz szegmentálását, illetve annotálását a Connexor vállalat Machinese nevű alkalmazásával végeztük, mivel ennek a programnak létezik egy demó változata, amely mindenki számára hozzáférhető és ingyenes. Ez egy online demó, amely a http://www.connexor.eu/technology/machinese/demo/ oldalról érhető el. A szövegablakba be kell írni az adott szöveget, meg kell adni a nyelvet, majd az egy HTML-fájlban visszaadja az adott szöveg mondatokra és tokenekre bontott változatát, amelyben minden token mellett annak szótövesített alakja és nyelvtani kategóriája is látható. A Machinese-ben két demó is rendelkezésünkre áll: az egyik csak egy POS-tagger (Machinese Phrase Tagger), a másik egy parser (Macinese Syntax), tehát szintaktikai elemzést is végrehajt. Annak ellenére, hogy a szintaktikai elemzést nem vesszük figyelembe, mi mégis az utóbbi demót fogjuk használni, mert ez az az alkalmazás, amely a legtöbb esetben egyértelműsít is, a POS-tagger esetében viszont azt tapasztaltuk, hogy ott sokszor sok lehetséges elemzési variánst megmutat.

A program működésének szemléltetése végett vegyük az alábbi példamondatot, amelynek szó szerinti fordítását is megadjuk, mert így ellenőrizhető le, hogy mennyire hatékony ez a szintaktikai elemző:

Nous	*voulons*	*que*	*tu*	*sois*	*heureux*	*de*	*tes*	*succès*
Mi	szeretnénk	hogy	te	legyél	boldog	miatt	a te	siker

'Azt szeretnénk, hogy örülj a sikereidnek.'

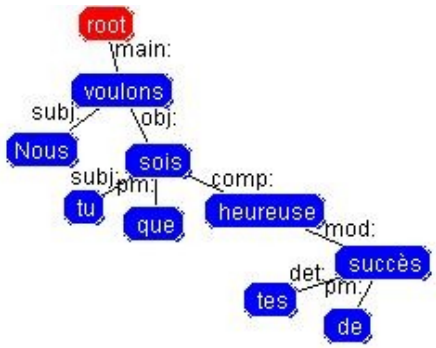
A szintaktikai elemző a fenti példamondatra a 7.1. táblázatban szereplő elemzést adta. Ezen keresztül mutatjuk be röviden a Machinese Syntax demójának működését, illetve annak jelölési rendszerét.

7.1. Táblázat: Példamondat elemzése a Machinese demójával

#	Text	Baseform	Syntactic relation	Syntax and morphology
1	Nous	Nous	subj:>2	&NH <Pers> PRON NOM PL1
2	voulons	vouloir	main:>0	&+FMA V IND PRES PL1
3	que	que	pm:>5	&PM> CS
4	tu	tu	subj:>5	&NH <Pers> PRON NOM SG2
5	sois	être	obj:>2	&+FMA V SUB PRES SG2
6	heureuse	heureux	comp:>5	&NH A FEM SG
7	de	de	pm:>9	&PM> PREP
8	tes	tu	det:>9	&DN> <Pers> <Poss> DET MSC PL
9	succès	succès	mod:>6	&NH N MSC PL
10	.	.		
11	<s>	<s>		

Az első oszlop a mondatban szereplő szavak mondatbeli elhelyezkedésének számát jelöli, a második oszlop a szöveg adott tokeneit, a harmadik oszlop azok szótöveit, a negyedik a többi elemhez való szintaktikai viszonyt jelzi, míg az utolsó oszlop a szófaji címkét jelöli. A szavak számozására elsősorban a negyedik oszlop miatt van szükség, ahol a szintaktikai viszonyoknál a szám azt jelzi, hogy a szintaktikai reláció hányas számú elemmel történik. A 4-es számú *tu* 'te' például az 5-ös számú *sois* ige alanya (subj:>5). Az ötödik oszlop elég részletes leírást ad az adott szövegszó adott környezetben vett szófajáról. Például az 5. sorban lévő ige, azaz az *être* 'lenni' szótővel rendelkező *sois*, egy ragozott, azaz finit (+F) főige (M), aktív használatban (A); ezen felül ige (V) kötőmódban (SUB), jelen időben (PRES) és egyes szám (SG) második (2) személyben. A 9. sorban lévő *succès* 'siker' a 6-os számú *heureuse* 'boldog' bővítménye (mod:>6), főnévi fej (NH) és egyben főnév (N), ami hímnemű (MSC) és többes számban áll (PL). A 11-es sorban szereplő <s> a mondat végét jelzi.

Érdekességképpen említhetjük, hogy a Machinese szintaktikai fát is tud rajzolni az adott szövegről, ez a 7.2. ábrán látható:



7.2. ábra: Példamondat elemzése függőségi fával (Machinese)

A 7.2. ábra alapján látható, hogy a Machinese nem a hagyományos értelemben vett generatív fát adja, hanem egy függőségi fát, ahol van egy gyökércsomópont, ahová a főmondat igéjét tesszük. Ez egy gráf, amelynek élei két végén a kapcsolatban álló elemek állnak, és minden élre igaz, hogy az alul elhelyezkedő csomópont a felette lévőnek a dependense.

A tokenizálást és szófaji egyértelműsítést nem úgy végeztük el, hogy a szövegeket bemásoltuk erre a weboldalra, a kimenetet pedig lementettük egy fájlba. Ez a lehetőség már csak azért sem merült fel, mert ez a demó csak bizonyos mennyiségű karaktersorozatot tud elemezni egyszerre, minden egyes bekezdés kézzel történő bemásolása, majd annak kézi mentése sok időbe került volna, így erre írtunk egy futtatható osztályt. Ez az osztály a szöveget bekezdésenként felküldi a Machinese szerverre, és a visszajövő HTML-oldalt elmenti. A tokenizált és szófajilag is elemzett szabadalmi leírásokat külön elmentettük. A TE további lépéseit ezen a korpuszon hajtjuk végre.

7.1.2. Tokenek kezelése

Az előző alfejezetben (7.1.1.) végrehajtott előfeldolgozás után egy HTML-oldalt kaptunk, ami az adott szabadalmi leírás szótövesített változatát tartalmazza táblázatként. Ez az oldal azonban nem alkalmas közvetlen feldolgozásra, ugyanis a tokeneken többször végig kell menni, másrészt sok felesleges információt tartalmaz.

Az annotálást tartalmazó dokumentumot először beolvastattuk, majd azok tokeneit egy rendezett listába helyeztük el. E dinamikus adattípus (Cormen és mtsai 2003) választása azért előnyös, mert a későbbiekben kiszűrt elemeket könnyen tudjuk belőle törölni. Ezen kívül a lista elemein történő iteráció a Java nyelvben könnyen megoldható, mivel az elemekre azok indexével hivatkozhatunk (Angster 2004). A lista adattípus főbb

műveletei közül az adott elem törlésére és módosítására a lista iterálására használatos függvényeket használjuk, amelyek a Java már előre beépített függvényei közé tartoznak.

A lista tokeneket tartalmaz, amelyhez létrehoztunk egy példányosítható Token osztályt. A Token osztály egy adott tokenről négy információt tartalmaz: a token szövegbeli előfordulási alakját, szótövét, annak morfoszintaktikai címkéjét és ez utóbbinak egyszerűsített változatát. A szintaktikai címkék egyszerűsítését azért végeztük el, mert nem minden morfoszintaktikai információra van szükségünk a TE során: egy igénél nem teszünk különbséget a között, hogy az kijelentő vagy kötőmódban van-e, valamint az sem érdekes, hogy az adott szövegben éppen hanyadik szám hanyadik személyben fordul elő. Ilyen esetben azokat egy egyszerű V címkével láttuk el. Az egyszerűsített morfoszintaktikai változókat a 7.2. táblázatban mutatjuk meg.

7.2. táblázat: A használt szintaktikai kódok

Címke	Jelentés
N	főnév
PREP2	*pour*, *sans* prepozíció
PREP	többi prepozíció
A	melléknév
ADV	határozószó
INF	főnévi igenév
V	ige
PRON	névmás
DET	determináns
NUM	számnév
CS	alárendelő kötőszó
CC	mellérendelő kötőszó
INTERJ	indulatszó
SIGN	bármilyen jel, kivéve / és -
SIGN2	- és /
X	általunk használt határoló

A 7.2. táblázatból látható, hogy voltak olyan morfoszintaktikai kódok, amelyeket kettéválasztottunk: erre például szolgálhat a kétféle prepozíció, ugyanis a *pour* sokkal gyakrabban szerepel igei bővítményként, mint terminus részeként, így azt a PREP2 kategóriába különítettük el. Bevezettük még a SIGN2 kódot is, mert ez gyakran lehet terminust alkotó tag, például az *analyse coûts-bénéfices* 'költség-haszon elemzés' esetében. Az X esetére a 7.2.2. fejezetben térünk vissza.

7.2. Szűrők

A tokenek listájának létrehozása után kezdtük el azok szűrését. Az első tapasztalatok alapján a két legfontosabb szűrendő elem a tulajdonnevek és azon köznyelvi fordulatok szűrése, amelyek biztosan nem lehetnek terminusok vagy azok részei.

7.2.1. Tulajdonnevek szűrése

A tulajdonnevek szűrésére nem készítettünk saját modult, hanem egy külső alkalmazást importálunk be a saját programba. A névelemek felismerésének összetettségét már az is igazolja, hogy akár külön disszertációk (pl. Benajiba 2009, Nadeau 2007) témáját is alkotják, így egy ilyen modul elkészítése külön ehhez a terminológiakivonatoló alkalmazásához nem lenne gazdaságos. A névelem-felismerő alkalmazás kiválasztásánál elsősorban három szempontot vettünk figyelembe: legyen (1) ingyenesen elérhető, (2) nyílt forráskódú, és (3) francia nyelvre is alkalmazható. Sőt, opcionálisan azt is szerettük volna, hogy Java nyelven írt legyen, hogy a programkódba közvetlenül beilleszthető legyen. Így döntöttünk az *OpenCalais Web Service API*[25] mellett.

Az OpenCalais Web Service egy webes alkalmazás, azaz a futásához internetkapcsolatra van szükség. A szerverre felküldött szöveget ez az alkalmazás automatikusan annotálja szemantikai metaadatokkal: jelöli – többek között – a benne előforduló személy-, cég-, és helyszínneveket. Az alkalmazás ingyenes, de használatához regisztrációs kód igénylése szükséges.

Az OpenCalais Java Eclipse fejlesztőkörnyezetbe történő importálásához Adjiman (2009) nyújtott segítséget, az általa publikált videó segítségével. Az általa mutatott eljárásnak köszönhetően a programcsomag közvetlenül felkerült a forrásfájlok közé. Ezen kívül egy próbaalkalmazás megírásában is támpontot nyújtott.

Az alkalmazás használata előtt könnyedén le is tesztelhettük annak működését, mert az OpenCalais névelem-felismerő moduljához már készült egy Mozilla Firefox add-on (kiegészítő) is, a *ClearForest Gnosis* (2009). Ezt telepítve, a böngészőben közvetlenül látható az eredmény: minden megnyitott weboldal esetén a névelemeket aláhúzással jelöli, és az aláhúzás színe határozza meg, hogy milyen típusú névelem lett kijelölve. Erre példa a 7.3. ábra, ami a ClearForest Gnosis futását szemlélteti egy angol nyelvű szövegen.

[25] http://www.opencalais.com/calaisAPI

7.3. ábra: ClearForest Gnosis futási példa[26]

A 7.3. ábrán jól látható, hogy a különböző típusú névelemek különböző színekkel vannak aláhúzva: a piros szervezetnév, a narancssárga földrajzi név, a zöld személy- vagy terméknév, a sárga kulcsfontosságú ipari terminusokat jelöl.

7.2.2. Konnektívumok és egyéb szókapcsolatok szűrése

A program első futása után derült ki, hogy szükséges olyan szókapcsolatokat is szűrni, amelyek nem lehetnek terminusok részei. Ezek a szókapcsolatok lehetnek főnévi csoportok vagy olyan melléknevek vagy határozószók, amelyek biztosan nem lehetnek terminusok részei. A konnektívumok is ide tartoznak, amelyek szorosan nem kötődnek a szaknyelvhez, hanem elsősorban a szöveg strukturáltságát, kohézióját biztosítják.

Ehhez a feladathoz először letöltöttünk egy olyan HTML-oldalt, amely ilyen kifejezéseket tartalmaz[27]. Ezt átalakítottuk szöveges formátumú fájllá, amelyben minden egyes sor egy konnektívumnak felel meg, és amit ezután kézzel ellenőriztük. Ezután ezeket kézzel szótövesítettük, és így mentettük el a szöveges fájlt. A TE során ezeket a

[26] https://addons.mozilla.org/img/uploads/previews/full/11/11978.png?modified=0
[27] http://www.colvir.net/prof/michel.durand/marqueurs.html

konnektívumokat láthatatlanná kellett tenni a mintaillesztési folyamatban. Így az *en réalité* 'valójában', *en effet* 'ugyanis', *par exemple* 'például' tokeneket X szófaji kódú elemre cseréltük, ami által a mintaillesztés során már ezek sem főnévi sem melléknévi elemei nem kerülhettek be a terminusjelölt-listába, mert például az *exemple* 'példa' szó a *par exemple* szókapcsolatban elvesztette főnévi címkéjét. A szótövesítés minimális erőfeszítést igényelt, és gyakran szabályos kifejezéssel is megoldható volt. Az *au contraire* 'épp ellenkezőleg' így *à le contraire*-ként, az *en d'autres termes* 'más szóval' pedig *en de autre terme*-ként került be a listába.

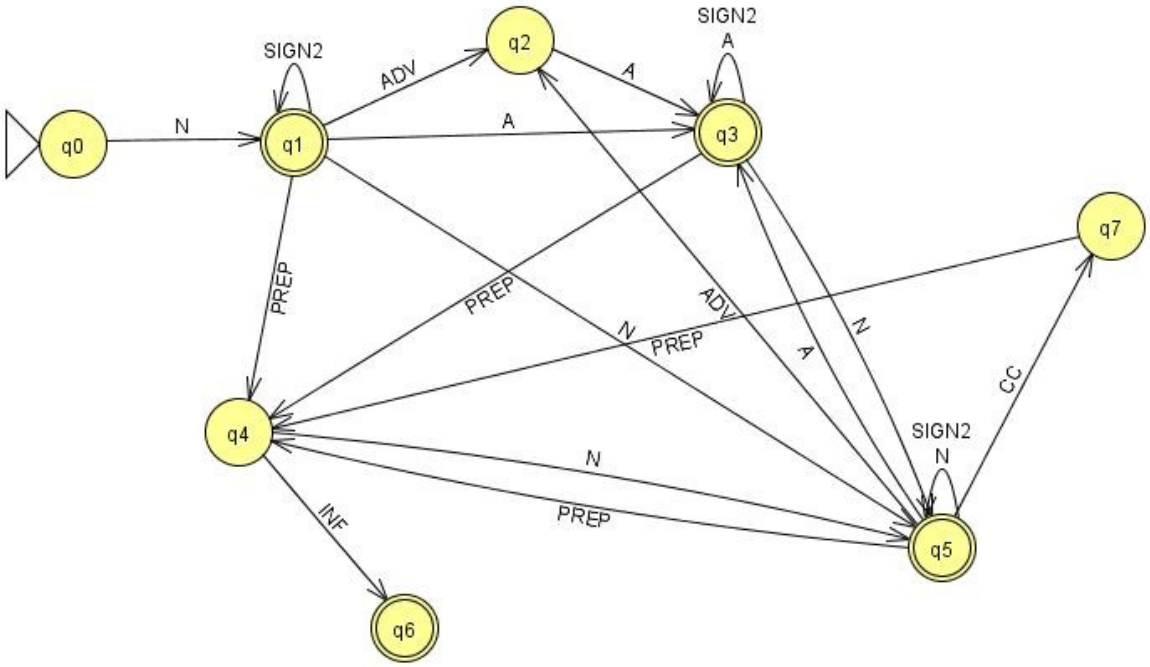
7.3. Automata létrehozása mintaillesztés céljából

A terminusjelölt-lista létrehozásához szükség volt egy véges állapotú automatára a reguláris kifejezések 5.3.1.2. alfejezetben felsorolt hátrányai miatt (pl. nehéz bővítés). A terminusokat azok belső szintaktikai összetétele alapján nyertük ki, így az automatának a tokenekről csak azok szintaktikai kategóriáját kellett ismernie. Az 5.3.1.2. fejezetben leírtuk, hogy a determinisztikus véges állapotú automatákat egy rendezett ötös ($Q, \Sigma, \delta, q_0, F$) ír le:

„Az $M = (Q, \Sigma, \delta, q_0, F)$ rendszert determinisztikus automatának nevezzük, ha:
1. Q egy nem üres, véges halmaz, az *állapotok halmaza*,
2. Σ egy ábécé, az *input ábécé*,
3. $q_0 \in Q$ a *kezdő állapot*,
4. $P \subseteq Q$ a *végállapotok halmaza*,
5. $\delta : Q \times \Sigma \rightarrow Q$ egy leképezés, az *átmenetfüggvény*" (Fülöp 2004)

A terminológiakivonatoló alkalmazásunknál az állapotok halmaza a konvenció szerinti q_0, q_1, q_2 … q_n számozást követi. Az alsó indexben lévő számok itt csak az adott állapot azonosítására szolgálnak: a lényeg, hogy minden állapotnak más azonosítója legyen. Ezen állapotok közül a q_0-t kijelöltük kezdőállapotnak. Mivel az automata csak szófaji kódokat olvas, így az input ábécét a szófaji címkék alkotják, azon belül is az általunk 7.1.2. pontban felsorolt egyszerűsített szófaji kategóriák (pl. N a főnév, A a melléknév, ADV a határozószó). Az átmenetek és a végállapotok halmazát úgy alakítottuk ki, hogy az automata a főnévi terminusoknak megfelelő mintákat kinyerje. Az automatát determinisztikusnak terveztük, mert ennek jobb a futási ideje és átláthatóbb is. Ez azt jelenti, hogy az automata akármelyik állapotban is van, mindig csak egy állapotba kerülhet, akármelyik input betűt olvassa.

A 7.1.2. pontban felsorolt egyszerűsített szófaji kategóriák és a 4.2. alfejezetben felsorolt minták alapján az alábbi véges állapotú automatát hoztuk létre:



7.4. ábra: A terminológiakivonatoláshoz használt automata

Ebben az automatában a q_0 a kezdőállapot, ezt jelöli a kör előtti háromszög. A q_1, q_3 , q_5 és q_6 a végállapotok, ezt dupla körrel jelöljük. Az automatát ellenőrizve látható, hogy determinisztikus, mert minden egyes állapotban igaz, hogy akármelyik betűt is olvassa, mindig legfeljebb csak egy állapotba mehet tovább. A q_5 állapotból például az A címke hatására a q_3 állapotba, az N címke hatására a q_5 állapotba, a CC címke hatására a q_7 állapotba léphet tovább. Az automata működéséhez vegyünk néhány példát. (i) Az N mintájú *serveur* 'szerver', (ii) az N A mintájú *nombre hexadécimal* 'hexadecimális szám', (iii) az N N SIGN2 N mintájú *analyse coûts-bénéfices* 'költség-haszon elemzés', és (iv) az N PREP N PREP N PREP N mintájú *système de gestion de base de données* 'adatbáziskezelő-rendszer'. Ezek átmenetei a következők:

(i) $q_0 \rightarrow q_1$

(ii) $q_0 \rightarrow q_1 \rightarrow q_3$

(iii) $q_0 \rightarrow q_1 \rightarrow q_5 \rightarrow q_5 \rightarrow q_5$

(iv) $q_0 \rightarrow q_1 \rightarrow q_4 \rightarrow q_5 \rightarrow q_4 \rightarrow q_5 \rightarrow q_4 \rightarrow q_5$

Mint látható, az automata minden főnévi terminushoz köthető mintát felismert, mert minden állapotnál tovább tudott lépni, és mindig végállapotba érkezett (q_1, q_3 vagy q_5).

A fent látható automatát a JFLAP nevű programmal hoztuk létre, és módosítottuk, ha szükséges volt. A JFLAP egy olyan ingyenes alkalmazás, amelynek segítségével könnyedén, egérkattintással bármilyen automatát (véges állapotú vagy veremautomata)

vagy Turing-gépet létrehozhatunk, majd azokat vizualizálhatjuk, és működésüket szimulálhatjuk egy input szövegre. A program a Java Swing grafikus felületet használja, és számunkra egy JAR csomagban tölthető le regisztráció után. A kézzel létrehozott grafikon egy XML-fájlba menti .jff kiterjesztéssel, így az automata akár a program megnyitása nélkül is szerkeszthető (*JFLAP*). Az XML-formátum legnagyobb erénye abban rejlik, hogy a kézzel létrehozott automatát könnyedén beolvastathatjuk bármilyen egyszerű programkóddal.

Az automata feldolgozására írtunk egy osztályt, ami paraméterül azt az XML-formátumú fájlt várja, amelyik az automata JFLAP formátumát tartalmazza. A hatékonyabb mintaillesztés érdekében az automatából egy kétdimenziós tömböt hoztunk létre, amelynek segítségével az átmenetek táblázatos formában tárolhatók. A táblázatos megoldás sokkal hatékonyabb, mint ha az automatát egy gráf adattípusban tárolnánk (Cormen és mtsai 2003). A kétdimenziós tömb (vagy mátrix) első értéke az adott átmenet kiinduló állapotát tartalmazza, a második érték pedig a beolvasott szófaji címke kódját. A mátrix azt mutatja, hogy az első paraméterben lévő állapotból a második paraméterben megkapott szófaji címkékkel melyik állapotba kerül az automata. A fenti automatából létrehozott mátrixot a 7.3. táblázat mutatja, ahol a *-gal jelölt állapotok végállapotok, és ahol a -1-gyel jelölt állapotokban az automata megáll, és a következő input szófajkarakter-sorozat beolvasásakor a 0 állapotból újraindul.

7.3. táblázat: Véges állapotú automatából létrehozott mátrix

	N	**PREP**	**PREP2**	**A**	**ADV**	**INF**	**V**	**PRON**	**DET**	**NUM**	**CS**	**CC**	**INTERJ**	**SIGN**	**SIGN2**	**X**
0	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
1*	5	4	-1	3	2	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1
2	-1	-1	-1	3	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
3*	5	4	-1	3	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	3	-1
4	5	-1	-1	-1	-1	6	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
5*	5	4	-1	3	2	-1	-1	-1	-1	-1	-1	7	-1	-1	5	-1
6*	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
7	-1	4	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1

A táblázatból jól látható, hogy a véges állapotú automata (1) determinisztikus, mert minden állapotból egy adott szófaj olvasása után mindig csak egy állapotba mehet tovább; valamint (2) teljes, mert minden állapotban valamelyik állapotba továbblép, akármelyik szófajt kapja inputként. A (2) megvalósításához azon esetekben, amikor egy adott állapotból nem vezet egy adott címkével rendelkező él, nem a megszokott „csapda" állapotot hozzuk létre, hanem ekkor ezen esetben a q_0 állapotba irányítjuk az automatát, tehát az automata a kezdőállapotból folytatja tovább a mintaillesztést.

7.4. Minta illesztése

Véges állapotú automata segítségével több módszerrel lehet mintát illeszteni. Az egyik ilyen a NooJ nevű alkalmazás, amely egy természetesnyelv-feldolgozási célokat szolgáló nyelvi eszköz. A NooJ akár nagy terjedelmű korpuszok feldolgozására is használható: leggyakoribb felhasználási területe morfoszintaktikai minták, állandósult és félig állandósult szókapcsolatok és konkordanciák indexelése (*NooJ Manual*). A NooJ felhasználói általában véges állapotú automaták használatával jelölnek ki a megadott morfoszintaktikai mintákra illeszkedő szövegrészeket az inputban megadott korpuszokban.

A Magyar Számítógépes Nyelvészeti Konferencia köteteit böngészve látható, hogy a nyelvészek, különösen a Magyar Tudományos Akadémia Nyelvtudományi Intézetében elég gyakran használják ezt az alkalmazást különböző, szabályszerűségeket követő karaktersorozatok kinyerésére. Erre példa lehet a bibliográfiai hivatkozások kinyerése (Váradi és mtsai 2010), vagy érzelmek, kognitív állapotok felismerése mintákkal (Szalai és mtsai 2009).

Minthogy esetünkben a program többi része Java nyelven íródott, így amellett döntöttünk, hogy a mintaillesztés is Java nyelven történjen meg. Erre szolgál az a modul, amelynek bemenete a kétdimenziós tömbbé alakított automata, valamint az adott szöveg morfoszintaktikai kódokkal ellátott tokenjei egy listában. Az alkalmazás futása közben szintén egy listában tárolja a terminusok kezdő és végpontját, valamint azok gyakoriságát.

A mintaillesztés algoritmusának egyszerűsített pszeudokódja a következő:

```
(int) akt_állapot = elozo_allapot = 0;
(boolean) NP_nyitva=false; //még nem kezdődött el egyetlen NP sem
Amíg tart a szöveg
        elozo_allapot = akt_allapot;
        akt_allapot = az átmenet utáni állapot;
        beolvassuk a következő szó szófaji kódját
        ha az automata tartalmaz átmenetet másik állapotba ennek hatására
                NP_nyitva = true;
                Feljegyezzük az NP elejének helyét;
```

```
Amíg NP_nyitva == true
        Ha a mintaillesztés során bármikor végállapotba kerülünk,
        annak feljegyezzük a helyét
        elozo_allapot = akt_allapot;
        akt_allapot = az átmenet utáni állapot;
        Ha akt_allapot==0, akkor már nem tudtuk tovább
        illeszteni a mintákat
                Ha végállapotba érkeztünk, az aktuális hely lesz az
                NP vége
                Ha nem, akkor a legutóbbi végállapothoz
                kapcsolódó hely lesz a végállapot, és onnan
                folytatjuk tovább a mintaillesztést
                Ha nem volt eddig végállapot, akkor a kiindulási
                helyet eggyel növeljük, és onnan folytatjuk a
                mintaillesztést
        NP_nyitva = false;
növeljük a számlálót
```

A mintaillesztés során az alkalmazás mindig ellenőrzi, hogy olyan mintát illesztett-e, amely tartalmaz-e mellérendelő kötőszót, azaz CC szintaktikai címkével rendelkező elemet (pl. *et* 'és'). Ha ilyet tartalmaz, akkor azt legtöbbször két külön terminusra kellett szétbontani. Ilyen összetétel például *interface graphique de surveillance et de diagnostic* 'felügyeleti és diagnosztikai grafikus interfész' vagy *message de contrôle ou d'erreur* 'ellenőrző vagy hibaüzenet'. Ezen esetekben az adott mintára illeszkedő karaktersorozatot két különböző karaktersorozatra kellett bontani. Ekkor az első karaktersorozat a kötőszó előtti részig tart, az első példánál ez az *interface graphique de surveillance*. A második részhez először vettük a kötőszó utáni első elemet (amely ezen esetekben prepozíció), feljegyeztük annak morfoszintaktikai kódját, és megkerestük a kötőszótól visszafelé az első olyan elemet, amelynek ugyanez a kódja, ez az első példánál az első *de*. Ekkor a kötőszó utáni részt ezen pozícióba másoltuk, így kaptuk az *interface graphique de diagnostic* terminusjelöltet. A morfoszintaktikai kód ellenőrzésére azért van szükség, mert nem biztos, hogy az prepozíció. A *document vidéo et audio* 'audió és videó dokumentum' esetében például a kötőszó után melléknév található, így a kötőszó előtti melléknevet kellett megkeresni, ami az *audio*, így a *document vidéo* 'videó dokumentum' és *document audio* 'audió dokumentum' terminusokat kaptuk.

7.5. Statisztikai módszerek

A jelen alfejezet ismerteti a TE során szűrésre alkalmazott statisztikai módszereket. A választott módszerek mind különböző szempontok alapján vizsgálják a terminusjelölteket. A *termhood*-értékek (7.5.1.) a terminusjelöltek szaknyelvhez való kapcsolódását mérik, a *unithood*-mértékek (7.5.2.) azok egybetartozását, valamint a terminusjelöltek

szövegkörnyezetét figyelembe vevő értékek (7.5.3.). A három érték egy értékké történő egyesítését a 7.5.4. alfejezet mutatja be.

7.5.1. *Termhood*-érték kiszámítása

A *termhood* statisztikai módszerek azt írják le, hogy egy adott karaktersorozat mennyire kapcsolódik szorosan a szaknyelvhez és a köznyelvhez. Mivel itt elsősorban a szaknyelvi és a köznyelvi előfordulásokat vettük alapul, ezért elengedhetetlen, hogy a köznyelvi gyakoriságról is információkkal rendelkezzünk. A köznyelvi gyakoriság kiszámításához az interneten használatos keresőmotorok mellett döntöttünk, amelyek nemcsak egy adott begépelt karaktersorozatot tartalmazó releváns weblapokat adják vissza, hanem azon releváns weblapok megközelítő számát is, amely azt tartalmazza. Így az adott terminusjelöltekről már azt is tudtuk, hogy azok hány dokumentumban szerepeltek a világhálón. A keresési eredményeket szűrni kell, mert az interneten feltehetőleg rengeteg szakszöveg is található, ebből adódóan a keresés eredményeként kapott szám nem feltétlenül tükrözi az adott terminusjelölt köznyelvi szövegekben való gyakoriságát.

7.5.1.1. A világhálón történő keresés megvalósítása

Ahhoz, hogy minden egyes terminusjelöltnél megállapítsuk, hogy hányszor fordul elő köznyelvi szövegben, egy internetes keresőmotorra volt szükségünk. Mivel az interneten több, közel hasonló paramétereket váró motor létezik, ezért annak megfelelően választottunk, hogy számunkra melyik adta a legjobb keresési eredményt. Mivel többszavas terminusokat kerestünk, (1) fontos, hogy azokat a szavakat egymás után keresse, ne csak azt nézze, hogy az a három szó megtalálható-e az adott szövegben vagy sem. Mivel a keresés során szótövesített alakban kerestünk, (2) elengedhetetlen, hogy a keresőmotor ne csak a megadott keresőszavakat, hanem azok ragozott alakjait is képes legyen felismerni. Mivel a terminológiakivonatolót teljesen automatikusra terveztük, ezért (3) nélkülözhetetlen, hogy a keresőmotort a programkódon belül meg lehessen hívni, és annak eredményeit ki lehessen automatikusan nyerni a szervertől visszakapott weboldal szövegéből.

A legtöbb felhasználó által alapértelmezetten használt Google keresőmotorja a fent megadott feltételek közül közvetlenül egyiket sem tudja megvalósítani. A keresést ugyan szótövesített alakkal végre lehet hajtani, de azt már nem lehet megadni, hogy a szótövesített alakokat egymás után keresse a szövegben. A keresőszavak elé tett + jellel

megadható, hogy az a szó mindenképpen legyen a találatok között, de az egymás egymásmellettiségre nincs megfelelő operátor. Egyetlen lehetőség az idézőjelek közötti keresés, amelyben pontos karaktersorozatokra kereshetünk rá. Ekkor viszont csak és kizárólag erre a kifejezésre keres, nem veszi figyelembe azok morfológiai variánsait. Ha például a *système d'exécution* 'operációs rendszer' szótövesített alakjával (+système +de +exécution) keresünk, akkor olyan oldalakat is visszaad a rendszer ahol ez a szócsoport többes számban előfordul, de sajnos olyan oldalakat is, ahol ez a három szó egymástól nagyon távol áll, és lehet, hogy nem is alkotnak egybefüggő szókapcsolatot. Ha a másik módszerrel próbálkozunk (″système de exécution″), akkor csak olyan weblapokat kapunk eredményül, amelyben pontosan így szerepel ez a három szó egymás után. Ezen a tényen kicsit változtatna, ha ezt a főnévi csoportot nem szótövesítve írnánk be („système d'exécution"), mert ekkor a többes számú alakok vesznének el a keresés során („systèmes d'exécution"). Mivel a Google-t feltehetőleg sokan próbálják automatikus keresési célokra használni, ezért a vállalat már megszüntette korábbi alkalmazását, a *Google SOAP Search API*-t, amit Java nyelvű környezetbe is be lehetett építeni, majd abból automatikus keresést végrehajtani. Ráadásul a Google meghívásakor, ha a visszaadott találati oldal HTML-forrását próbálnánk megtekinteni, akkor ott már nem olvasható a konkrét találati szám, csak függvények nevei, ezért ebben az esetben egy másik keresőmotort kellett alkalmazni.

A választásunk az ExaLead vállalat keresőmotorjára esett, ami mindhárom feltételnek eleget tett. A különböző szóalakokra történő keresésre a keresőszó után megadott * karakter szolgált (pl. a *système* és *systèmes* megkeresésére a *système** keresőszó). A szavak egymásmellettiségének biztosítására a NEXT operátor alkalmas, például az *interface graphique* esetén *interface NEXT graphique* kifejezést kellett alkalmaznunk. A nyelv megadására pedig a *–language* paraméter szolgál, amelynek értékét FR-re állítottuk. Az ExaLead esetében a találati oldalról a találatok száma az oldal HTML-forrásából könnyen kinyerhető.

Azért, hogy a keresőmotor ne keressen bárhol az interneten, meg kellett adnunk egy olyan weblap nevét, amely elsősorban köznyelvi szövegeket tartalmaz. Sok más tanulmányhoz hasonlóan egy köznyelvi újságot választottunk referenciakorpuszként (pl. Drouin 2003, Gelboukh és mtsai 2010), a *Le Figaró*t. Igaz ugyan, hogy egy napilapból nem biztos, hogy hiányoznak a terminusok, de a fent felsorolt tanulmányok eredményei mind arra következtetnek, hogy ez egy megfelelő módszer lehet. Bizonyos terminusok – a TE számára hátrányos – köznyelvi elterjedéséhez elég annak mindennapossá válása (pl.

mobiltelefon) vagy egy olyan esemény, amelyről gyakran cikkeznek: elég csak a 2010-es évre gondolnunk, ahol a napilapokban korábban nem olvasott *vörösiszap*, illetve *zagytározó* terminusok hirtelen százszámra felbukkantak a különböző médiumokban. Ezen eseteket, természetesen, nem tudtuk kiszűrni, de ezen szavak még mindig ritkábban fordulnak elő, mint a terminusként kevésbé értelmezhető *probléma* vagy *baleset* szavak.

A keresési feltétel megadásakor tehát megadtuk azt is, hogy csak ennek a webújságnak az oldalain keressen: ezt az inurl:weblapnév segítségével tehettük meg. A *reconnaissance de données biométriques* 'biometrikusadat-felismerés' terminust az ExaLead keresőmotorral a francia nyelvű, *Le Figaró*n található oldalak közül az alábbi kifejezéssel kereshetjük:

inurl:lefigaro.fr language:fr ((+reconnaissance* NEXT +de*) NEXT +donnée*) NEXT +biométrique*

Minthogy ezt az internetes keresőmotort emberi lekérdezésekre tervezték, ezért a forgalmat figyelő szerver nem engedélyezett több lekérdezést egymás után. Ezért ezen lekérdezésekhez beépítettünk egy időfüggvényt is, amely minden egyes, az Exalead szerveréhez továbbított kérés esetén körülbelül két percet vár, ezzel is azt az érzést keltve számára, hogy a keresést egy ember végzi, nem pedig egy számítógép.

Annak elkerülésére, hogy a program minden egyes futásakor az összes terminusjelöltre ne kérdezze le a szervertől azok előfordulási gyakoriságát, ezeket a karaktersorozatokat egy fájlba mentettük azok előfordulási számával. Ezzel jelentős időt spóroltunk meg, hiszen már öt darab szabadalmi leírásából 1200 terminusjelöltet nyertünk ki, amelyeknél ha mindig két percet várnánk azok gyakoriságának lekérdezéséhez, az 2400 percet, azaz 40 órát venne igénybe. A program ezért mielőtt lekérdezné egy terminusjelölt gyakoriságát, megnézte, hogy az szerepel-e már ebben a fájlban, és ha igen, akkor azt az értéket veszi alapul.

7.5.1.2. *Weirdness*-érték kiszámítása

A *termhood*-értékek közül a *weirdness*t választottuk, amelynek értéke egy adott szónál az alábbi képlet alapján számolható ki (Ahmad és mtsai 1999):

$$weirdness(w) = \frac{\frac{f_s(w)}{t_s(w)}}{\frac{f_g(w)}{t_g(w)}}$$

A *w* a vizsgált szó, $f_s(w)$ a *w* előfordulási száma a szakszövegben, $f_g(w)$ a *w* előfordulási száma az általános korpuszban, $t_s(w)$ a szakszöveg összes tokeneinek a száma, míg $t_g(w)$ az általános korpusz tokeneinek száma. Azon szavak esetében, amelyek mind az általános, mind a szakszövegben gyakran és ugyanolyan aránnyal fordulnak elő, ott ezek értéke 1 körüli, míg a terminusoké, amelyek jobban meghatároznak egy szakszöveget, ennél nagyobb.

Mint ahogy az 5.4.1.3. fejezetben kifejtettük, ezt a módszert elsősorban nem TE-célokra fejlesztették ki, de arra is használható. A választásunkat az indokolja, hogy nem rendelkeztünk konkrét általános nyelvi korpusszal, amelyben tudnánk, hogy az egyes dokumentumok hány tokennel rendelkeznek, és ez a képlet az, amelyből ezt ki lehet hagyni. A $t_g(w)$ az általános korpusz tokenjeinek száma ugyanis nem ismert, csak az, hogy az adott karaktersorozat hány weblapon szerepel, de mivel ez az összes terminusjelölt esetén hiányzik, ezért a különböző karaktersorozatok esetén az arány ugyanúgy megmarad. A fenti képletet átalakítva:

$$weirdness(w) = \frac{f_s(w)}{t_s(w) \cdot f_g(w)} \cdot c$$

A képlet ugyanaz, mint az előző, csak a hiányzó értéket egy *c* konstansra cseréltük, amelynek értékét a konkrét tokenszám hiányában a dokumentumok számában határoztuk meg. A dokumentumok számának megadása abból a szempontból is előnyös, hogy az internetes keresés során visszakapott találatok száma nem azt mutatja meg, hogy az adott szó a megadott tartományban hányszor szerepel, hanem azt, hogy hány darab dokumentumban. Így ha csak a dokumentumok száma számít előfordulásnak, akkor az $\frac{f_s(w)}{t_s(w)}$ aránypár esetében is a nevezőben ezt az értéket kell megadni. A dokumentumok számának lekérdezéséhez az ExaLead keresőjébe csak annyit írtunk, hogy *inurl:lefigaro.fr*, így csak azon dokumentumok megközelítő számát kaptuk meg, amelyek ezen a weblapon találhatók. Erre az értékre 687.717 találatot kaptunk, de mivel ez valószínűleg becslés, így ezt 700.000-re kerekítettük.

Ha egy terminusjelölt az interneten egyáltalán nem fordult elő, akkor a képletben az $f_g(w)$ értéke 0, hiszen ez adja meg annak előfordulási számát. Ekkor viszont a képletben nullával kellene osztani, ami lehetetlen. Ha az $f_g(w)$ kifejezés értéke 0, akkor az a *Le Figaro* oldalán elvileg nem található, így annak nagy értéket kell kapnia, ezért ezen esetekben maximális *weirdness*-értéket adtunk. Mivel a *weirdness*-értékek

reprezentálásához és tárolásához a float lebegőpontos formátumot alkalmaztuk, így ide csak ilyen értéket adhatunk meg. Végtelen érték nincs, ezért ekkor a float tartomány legnagyobb elemét adtuk meg, amelyet a Java nyelvben a Float.maxValue konstans szolgáltatja. Fontos még az is, hogy ha a *weirdness*-érték egynél nagyobb, akkor az azt jelenti, hogy a fenti képletben a számláló nagyobb, mint a nevező, tehát az a terminusjelölt gyakrabban fordul elő a szakszövegben, mint a köznyelvben.

A 7.4. táblázatban egy-két terminusjelöltön mutatjuk be azok előfordulását a szabadalmakban és a *Le Figaro* oldalról. A táblázat sorait azok *weirdness*-értékei szerint rendeztük csökkenő sorba. A 7.4. táblázatban olyan elemek statisztikáit is megadjuk, amelyek nem terminusok (ezeket aláhúzással jelöltük), hogy látszódjon az is, hogy ez az érték nem minden esetben mérvadó.

7.4. táblázat: Példák különböző gyakoriságú terminusjelöltek weirdness*-értékeire azok gyakoriságának feltüntetésével*

Terminusjelölt	Fordítás	Exalead gyakoriság	Szabadalmi előfordulás	Weirdness
paquet IP	IP-csomag	0	1	∞
mot de code correcteur d'erreur	hibajavító kódszó	2	25	2305,0579
document multimédia	multimédia-dokumentum	3	20	1039,1154
donnée biométrique	biometrikus adat	44	33	138,3035
solution connue	ismert megoldás	2	1	77,9337
serveur de contenu multimédia	multimédia-szerver	13	5	59,94896
opération de maintenance	karbantartás	134	11	16,68
matrice	mátrix	226	11	7,5865
couche de communication	kommunikációs réteg	57	1	3,3508
circuit de commande	vezérlő áramkör	230	1	0,8834
invention	találmány	2370	9	0,7716
confidentialité	megbízhatóság	2851	8	0,5174
requête	lekérdezés	2332	4	0,4914
serveur	szerver	2622	6	0,4371
étape	szakasz	18413	13	0,1435
réseau	hálózat	70830	43	0,116
mise en oeuvre	üzembe helyezés	6995	5	0,1114
description	leírás	10159	5	0,1
avantage	előny	19083	1	0,01
vidéo	videó	284945	6	0,0033
homme	ember	125012	1	0,0020
fin	vég	193717	1	0,0008

A táblázatból jól látszik, hogy az alacsony *weirdness*-értékkel rendelkező elemek között van a legtöbb olyan elem, amely vagy nem terminus (pl. *fin, étape*), vagy gyakran használatos nem terminus jelentésben (pl. *vidéo, description*). A magasabb értékekkel rendelkező elemek között is akad azonban olyan jelölt, amely biztos nem terminus, például a *solution connue* 'ismert megoldás'. Ez mind azt mutatja, hogy egyedül ezen érték használata nem lehet elegendő a statisztikai szűrés folyamán. Ezt a tényt támasztja alá az is, hogy a keresés során a jelentést nem vettük figyelembe, így például a köznyelvben gyakran használt *serveur* szó inkább 'felszolgáló', mint 'szerver' jelentésben fordul elő.

7.5.2. *Unithood*-érték kiszámítása

A *unithood*-érték egy terminusjelöltnél azt mutatja meg, hogy annak elemei mennyire koherensek egymással, azaz az adott terminusjelölt tényleg egybetartozik-e, vagy egy nagyobb egység része, vagy csak annak egy része terminus. Az 5. fejezetben felsorolt mértékek közül a két legsokrétűbb a C/NC, valamint az IR/CF érték. Mindkettőre jellemző, hogy az adott terminusjelölt környezetét figyelembe veszi, valamint azt is, hogy az adott jelölt inkább egy nagyobb terminusjelölt része-e vagy sem.

Az elemzésünk során *unithood*-értéknek a 5.4.2.2. fejezetben ismertetett C-értéket választottuk, amelynek formulája a következő:

$$C\text{-}value = \log_2 |a| \cdot f(a) \qquad \text{ha } a \text{ nem beágyazott}$$

$$C\text{-}value = \log_2 |a| \cdot f(a) - \frac{1}{P(T_a)} \sum_{b \in T_a} f(b) \qquad \text{egyébként}$$

Az *a* a vizsgált terminusjelölt, *f(a)* annak előfordulási száma, |a| a terminusjelölt hossza szavakban mérve, T_a azon terminusjelöltek halmaza, amely tartalmazza *a*-t, $P(T_a)$ pedig azon terminusok száma, amelyek tartalmazzák *a*-t és hosszabbak is nála.

Ezen érték kiszámításához tehát szükségünk volt minden egyes terminusjelöltnél (1) annak hosszára, (2) azon elemek számára, amelyek tartalmazzák azt, és hosszabbak is nála, valamint (3) ez utóbbiak előfordulási gyakoriságára. A C-érték kiszámítását akkor végeztük el, amikor a mintaillesztési folyamat már lezárult, azaz az összes terminusjelölt a rendelkezésünkre állt.

Elsőként a terminusjelölteket sorba rendeztük hosszuk, azaz a bennük található szavak száma szerint (két, azonos számú szóból álló terminusjelölt között nem tettünk különbséget). A rendezéshez az egyik legjobb futási idővel rendelkező algoritmust, a kupacrendezést alkalmaztuk (Cormen és mtsai 2003).

A C-érték kiszámításhoz az alkalmazás végiglépett az összes terminusjelöltön azok hossza szerint csökkenő sorrendben. A legnagyobb értéknél csak a hosszának kettes alapú logaritmusát szorozta annak előfordulási gyakoriságával. Egy háromszor előforduló nyolcszavas, maximális terminus C-értéke tehát $\log_2 8 \cdot 3$, azaz 9. Ha olyan elemhez értünk, amelynek hossza nem maximális, akkor megkerestük azon elemeket, amelyek ennél hosszabbak, majd megnéztük, hogy azok közül hány különböző tartalmazza, és azok hányszor fordulnak elő. Ha egy négyszer előforduló kétszavas terminust összesen 4 darab terminusjelölt tartalmaz 8 darab különböző előfordulással, annak C-értéke $\log_2 2 \cdot 4 - ¼8$, azaz 2; ha a nagyobb terminusok összesen 16-szor fordultak elő, akkor a C-érték nulla lenne, ha pedig ennél többször, akkor negatív szám. Egy terminusjelölt C-értéke minél nagyobb, annál valószínűbb, hogy nem egy nagyobb egység része, hanem önmagában is terminusjelölt. Ez az érték minél kisebb, annál több olyan elem található, amely tartalmazza azt, tehát akkor azt nem kell terminusjelöltnek megtartani. Az egyszavas terminusok C-értékét nem lehet kiszámítani, mert azon esetekben az első tag biztosan nulla értékű lesz, mert $\log_2 1$ értéke 0. Így ha már létezik legalább egy terminusjelölt, amely azt tartalmazza, akkor az érték negatív lesz, és ha az egyszavas terminus külön is gyakran szerepel, akkor is eleve negatív értéket kapunk, ami félrevezető lenne, így az egyszavas terminusoknál ez nem alkalmazható.

A 7.7. táblázatban különböző terminusjelöltek C-értékeit mutatjuk meg. Ügyeltünk arra, hogy ne csak terminusok szerepeljenek a listában, a nem terminusokat aláhúzással jelöltük.

7.7. táblázat: Példa C-értékekre

Terminusjelölt	Fordítás	Előfor-dulás	C-érték
mot de code correcteur de erreur	hibajavító kódszó	25	62,0391
ensemble de données biométriques	biometrikus adatok halmaza	26	44,2222
donnée biométrique	biometrikus adat	33	29,9189
paiement électronique	elektronikus fizetés	19	18,0
terminologie anglo-saxonne	angolszász terminológia	12	12,0
adresse internet	internetcím	7	7,0
document multimédia appartenant	(...-hoz) tartozó multimédia-dokumentum	2	3,1699
couche réseau	hálózati réteg	3	3,0
sens de rotation adopté	alkalmazott forgási irány	1	2,0
écran tactile	érintőképernyő	2	2,0
façon de naviguer	navigálás módja	1	1,5850
message de erreur	hibaüzenet	1	1,5850
message arp	arp-üzenet	1	0
mémoire tampon	puffer	1	-3,0
fonctionnement normal	normális működés	1	-10,0

A táblázatban látható, hogy mind az alacsonyabb, mind a magasabb értékeknél előfordul nem terminus elem. Ez azt mutatja, hogy ez a mérték sem mutatja meg egyértelműen, hogy egy adott szövegben melyek a terminusok.

7.5.3. Terminusok súlya

Ahhoz, hogy megállapítsuk azt, hogy egy adott terminusjelölt tényleg terminus-e, megvizsgáltuk a környezetében lévő tokeneket. Ha ugyanis olyan token előzte meg vagy követte, amely tipikusan terminusokkal áll együtt, akkor az is nagyobb valószínűséggel volt terminus. A francia nyelvben tipikusan terminust bevezető kifejezés az *est appelé* (’valaminek nevezett’), de ide tartoznak a determinánsok is, amelyek főnévi terminusokat vezethetnek be és azoknak nem részei.

Az 5. fejezetben két módszert is említettünk környezetszavak súlyozására: a C-NC érték kiszámítására szolgáló Weight valamint az IR/CR algoritmusnál alkalmazott LD (*left dependency rate*) és RD (*right dependency rate*). Mivel a C-NC értéket választottuk az elemzésnél, így a kontextus figyelésénél a Weight súlyozó értéket alkalmaztuk a határolóegységek súlyának meghatározásához. A Weight értékét kiszámító képlet a következő:

$$Weight(w) = 0{,}5 \cdot \left(\frac{t(w)}{n} + \frac{ft(w)}{f(w)} \right)$$

A képletben a *w* a vizsgálandó kontextustoken (pl. határozott névelő, vessző stb.), *n* azon biztos terminusok egyszeri előfordulási száma, amelyekre ez a vizsgálat kiterjed (például ha 100 biztos, különböző terminusjelöltet választunk tanulókorpusznak, akkor 100), *t(w)* azon esetek száma, ahol egy biztosnak jelölt terminus ezzel a *w* szóval áll legalább egyszer együtt (az előző példánál maradva ez legfeljebb 100 lehet), *ft(w)* azt mutatja meg, hogy a *w* szó összesen hányszor fordul elő terminusokkal együtt, *f(w)* pedig *w* korpuszbeli előfordulásainak száma.

A képletben az *ft(w)* érték mutatja azt, hogy ez egy tanulóalgoritmus, mert szükségünk van egy olyan korpuszra, amelyben jelölve vannak a terminusok, így kinyerhető azok környezete is. Azonban nem rendelkeztünk ilyen korpusszal, ezért más módszert kellett választanunk. Ezért alkottunk egy olyan listát, amely száz terminus szótövesített változatát tartalmazza, és azok közül is azokat, amelyek gyakran szerepelnek, valamint azokat, amelyek minden esetben terminusok lehetnek csak. Ezt a listát a későbbiekben csak a súlyértékek kiszámítására alkalmaztuk.

Ezen érték kiszámítására egy külön függvény szolgált, amely először beolvasta ezt a felsorolást tartalmazó fájlt, majd megnézte, hogy a kinyert terminusjelöltek közül a biztos terminusok környezetében milyen tokenek álltak. A környezetben lévő tokenek tárolása a mintaillesztési algoritmus során történt, így ebben a szakaszban nem kellett a szövegeken még egyszer végigmenni, csak ezen adatokat elő kellett venni. A környezetben előforduló tokenek közül nem mindegyik kapott értéket, hiszen lehetséges, hogy nem minden token fordult elő olyan terminusjelölt közelében, amely a tanulólistában szerepelt. Ekkor alapértelmezés szerint ez az érték nulla lett. A tokeneket a szótövük és azok szófaji címkéje alapján tároltuk, feltételezve, hogy ezen adatok egyértelműen azonosították az adott szót: nem mindegy például, hogy az *informatique* éppen ’informatika’ jelentésű főnévként vagy ’informatikai/számítógépes’ jelentésű melléknévként követi a terminusjelöltet, de az, hogy például az *être* ’lenni’ ige melyik alakjában (kötőmód vagy kijelentő mód) előzi meg vagy követi, az nem befolyásolja az eredményt. A 7.6. táblázatban soroljuk fel a legnagyobb súlyértékkel rendelkező tokeneket:

7.6. táblázat: A 15 leggyakoribb weight értékkel rendelkező token

	Token szótöve	Token szófaja	Weight
1	elles	DET	0,50980395
2	interfacer	V	0,50490195
3	présent	A	0,45490196
4	,	SIGN	0,4329732
5	concerner	V	0,3977591
6	de	PREP	0,37154862
7	coder	V	0,34313726
8	en tant que	PREP	0,33823532
9	la	DET	0,3233778
10	capter	V	0,304902
11	un	DET	0,291439
12	le	DET	0,2905085
13	.	SIGN	0,28751418
14	es	DET	0,27734777
15	les	DET	0,26827785

A nagyobb súlyérték azt jelenti, hogy az adott elem legtöbbször terminusokkal fordul elő. A táblázatból látható, hogy a determinánsok szerepeltek a legtöbbször a főnévi terminusok közelében. A leggyakoribb a DET címkéjű *elles* 'ők' volt, amely csak a szótövesítő nem megszokott jelölése a *leur(s)* 'az ő …-k' birtokos determinánsra. A második helyen lévő *interfacer* egy POS-taggelési hiba miatt került bele a listába, ugyanis a korpuszban egyszer szerepelt, akkor is *interface* 'interfész' főnévi jelentésben, de abban az egy esetben terminus követte, ami a *réseau* 'hálózat'. A *présent* 'jelen' melléknév úgy került a 3. helyre, hogy általában az *invention* 'találmány' szóval áll együtt szabadalmak leírásaiban, mint például *la présente invention* 'a jelen találmány'. A determinánsokon kívül melléknevek és igék, prepozíciók és központozási jelek is szerepelnek a listában.

Maynard és Ananiadou (2000) célja az volt, hogy a *weight*-értékek alapján egy olyan értéket hozzon létre, amely a környezetében lévő szavak súlya alapján egy súlyértéket ad magához a terminushoz is. A terminushoz rendelt súlyt ők *wei*nek jelölik, képlete a következő:

$$wei(a) = \sum_{b \in C_a} weight(b) + 1$$

Az *a* terminus súlyát tehát a környezetében lévő szavak súlyának összegéből kapjuk, amelyhez egyet adunk. Az egy hozzáadásához valószínűleg azért volt szükség, hogy a súly is hasonló mértékben számítson bele az általuk létrehozott kombinált C- és NC-értékhez, mint a C-érték. Ahogy az eddigi mértékeknél, a terminusok súlyértékeinek esetében is készítettünk egy táblázatot (7.8.), amely támpontot nyújt az érték eredményéről.

7.8. táblázat. Terminusok súlyértékei

étant ’lévén’	1
mp4	1
interface	1,15
millier ’ezernyi’	1,17
instant ’pillanat’	1,34
structure de matrice ’mátrixszerkezet’	1,4
réseau ’hálózat’	2,01
forme ’alak’	2,01

A táblázat értékeiből látható, hogy az aláhúzott nemterminusok valamint a nem aláhúzott terminusok egyaránt előfordulhatnak mind magas, mind alacsony súlyértékekkel.

7.5.4. Összevont érték

A súly-, a C- és *weirdness* értékek kiszámítása után azokból egy összevont mértéket kellett létrehoznunk, amelyre igaz az, hogy amennyiben egy adott terminusjelölt ennek a mértéknek egy adott értékét átlépi, akkor az nagy valószínűséggel terminus. Ehhez (1) először a három mértéket egy adott tartományra kellett leképezni (pl. 0 és 1 közé), hogy egymással kompatibilisek legyenek. Ezt követően (2) meg kellett határozni, hogy az egyes mértékek mennyire fajsúlyosak a terminuskivonatolás során (pl. a *weirdness* érték 60%-ban, az NC- és C-érték 20%-ban számítson bele az összevont értékbe). Végül (3) meg kellett állapítani egy küszöbértéket az összevont értékre dokumentumonként, amely felett a fedés és a pontosság együttes értéke, az F-érték, maximalizálható: ezeket a küszöbértékeket a 8. fejezetben ismertetjük.

7.5.4.1. A három mérték közös tartományba történő leképezése

A három mérték egységes tartományba való leképezéshez a 0 és 1 közötti irracionális[28] tartományt választottuk, mert így a valószínűségszámításban használatos függvényekkel is számolhatunk, ugyanis a valószínűségi értékek mindig a [0;1] tartományban találhatók, ahol 0 a lehetetlen esemény (pl. dobókockával hetest dobni) valószínűségét mutatja, az 1 pedig a biztos eseményét (pl. érme feldobása esetén az fej vagy írás lesz).

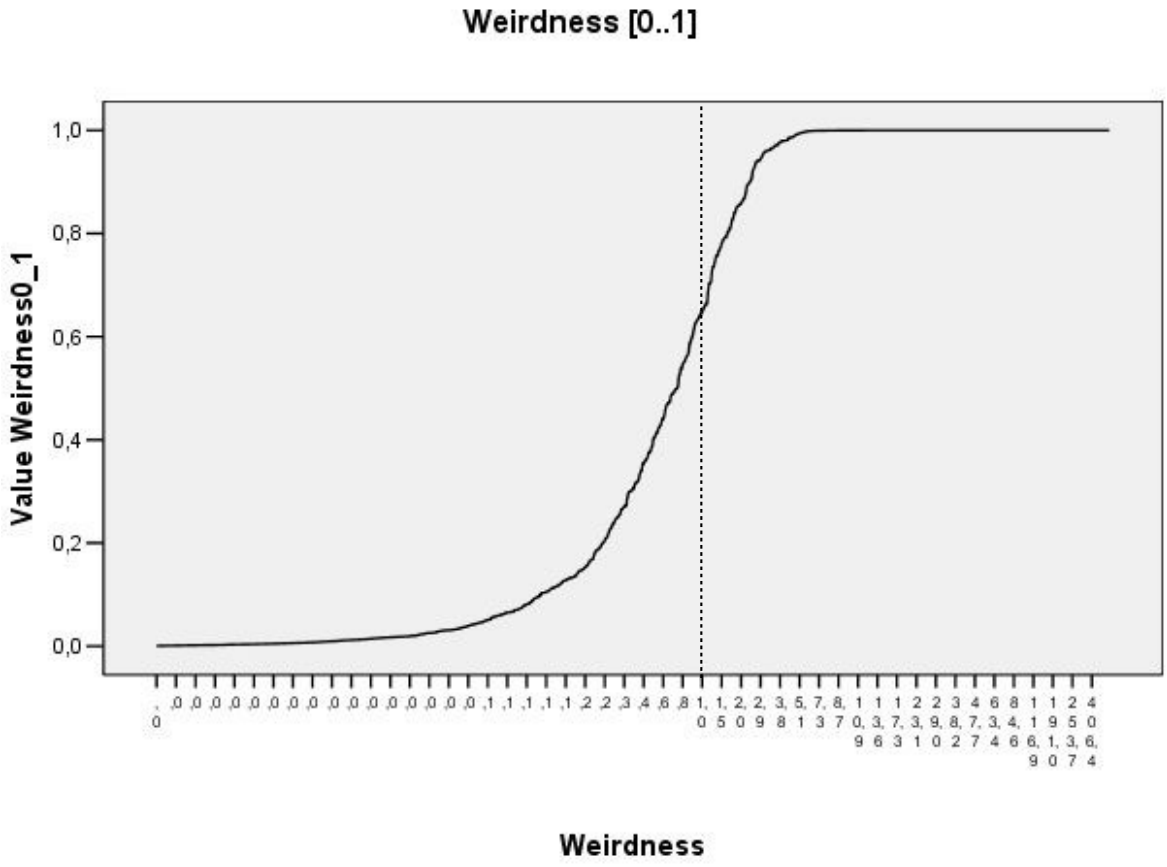
Mindhárom mérték esetében mind az átlagok és a szórások, mind a szélsőértékek és az eloszlások is különbözőek. Ezen értékek normalizálására szükség volt ezen adatok előzetes feldolgozására, az első öt elemzett korpusz adatainak felhasználásával, amelyek összesen 1357 értéket tartalmaztak.

[28] A *weirdness*-értéknél az exponenciális eloszlás valószínűségfüggvényével számoltunk, amelynek eredménye adhat irracionális értéket.

A *weirdness*-érték esetében a legkisebb érték a 0,00039, míg a legnagyobb 2305,058, a minták száma 979, mert a többi 378 (=1357-979) végtelen, azaz nem értelmezhető értéket adott. Végtelen értéket akkor kapott egy jelölt, ha az általános nyelvi korpuszban nem fordult elő, mert akkor a függvényben nullával kellett volna osztani, ami lehetetlen. A minta átlagértéke (μ=) 38,612 szórása (σ=) 149,625. Az átlagtól egy szórásnyi távolságra található elemek száma 920, tehát körülbelül 94%. A 0 és 1 közötti értékek száma 556, a maradék 403 ezen felüli érték, a medián 0,365. A mértékről az állapítható meg, hogy minél nagyobb az értéke, annál biztosabb, hogy a valószínűségi érték 1. Amiatt, hogy nagyjából ugyanannyi elem található a 0 és 1 közötti intervallumban, mint a felett, és amiatt, hogy a valószínűségi érték annál inkább egy, minél nagyobb a *weirdness*-érték, ezen értékek esetén az exponenciális eloszlás valószínűségfüggvényével számoltunk. Ehhez az ezen eloszláshoz tartozó eloszlásfüggvényt használtuk a valószínűségi értékek kiszámítására. De mivel arra voltunk kíváncsiak, hogy adott *weirdness*-értéknél milyen valószínűségi érték várható, ezért a képletben a $\xi \leq k$ helyett $\xi = k$ formulát írtuk.

$$F(x)=P(\xi=k)=1-e^{-\lambda x}$$

A λ paraméter a várható érték reciprokát jelöli. Mivel arra számítottunk, hogy az 1 feletti *weirdness*-érték esetén már többször fordul elő a terminusjelölt a szakszövegben, mint az általános nyelvi korpuszban, így az már majdnem biztos, hogy terminus. Azonban 1 felett nem szerettünk volna automatikusan p=1 valószínűségi értéket adni, így a fenti függvényben a λ paraméter értékét 1-nek választottuk, mert a várható értéknek 1-et feltételeztük, ezért λ = 1/1 = 1. Az alábbi grafikon mutatja a [0;1] tartományba leképezett értékeket az eredeti *weirdness*-értékek alapján:



7.5. ábra: Weirdness értékek leképezése a [0..1] tartományba

A C-érték legkisebb értéke -10,3023, legnagyobb 62,0391, a minták száma 822, mert a többi 535 egyszavas terminus volt, amelyekre ez a mérték nem alkalmazható. Az átlag 2,778, a medián 2, a szórás 4,4325. Az átlagtól egyszórásnyi távolságra található elemek száma 758, tehát körülbelül 92%. A mértéket nem volt érdemes egyenletes eloszlásúnak vagy exponenciális változónak tekinteni, mert az értékek nagy része azonos, és egy adott tartományban helyezkedik el. A legtöbbször az 1,5850 szerepelt 179 előfordulással, a perifériájában pedig a 0 szerepel 42 találattal, míg a 3,1699 43-szor. E két határ között 674 előfordulás található, amely az értékek kb. 82%-a. Így a C-értéket normál eloszlásúnak tekintettük, mert az értékek egy adott érték köré csoportosulnak, bár a három-szigma törvény, vagy másképp a 68-95-99.7 szabály (Pukelsheim 1994), azaz hogy az átlag körül egyszórásnyira az elemek 68%-a található, kétszórásnyira 95, háromszórásnyira pedig 99,7%, nem teljesül rá.

Annak ellenére, hogy a minta normál eloszlású, a normál eloszlásra használható eloszlásfüggvény nem alkalmazható a C-értékre, mert az az átlaghoz közeli értékekhez ad nagy értéket, viszont arra volt szükség, hogy minél nagyobb értékekhez minél nagyobb valószínűséget rendeljünk. Mivel az értékek nagyobb része az átlagtól egyszórásnyi

távolságon belül van, így a szélső értékeket 0-nak vagy 1-nek vettük. A fennmaradó elemeket pedig leképeztük a 0..1 tartományba egyszerű aránypárok segítségével.

A legkönnyebb dolgunk a súlyértékek esetén adódtak, amelyek várható értéke 1 és 2 közé esik. Azért legalább egy, mert a képletben a „+1” szerepel, amire csak azért van szükség, mert a C/NC-érték ebben az esetben hatékony. Mivel csak a terminusjelölt előtti és utáni értékeket vesszük alapul, ezért ez a mérték a környezetszavak súlyának maximumának legfeljebb kétszerese lehet. A súlyértékek átlaga 1,504, 0,20-as szórással, a minimum érték 1, a maximum 2,0196, az átlagtól egyszórásnyira az elemek 75%-a található. Mivel csak három darab 2 feletti érték volt, ezeket egynek vettük, a többi értékből pedig egyet levonva megkaptuk a 0 és 1 értékre levetített súlyértékeket.

7.6. Adatbázis létrehozása

A futási folyamat során kiszámolt értékeket, a kinyert terminusjelölteket egy adatbázisban tároltuk azért, hogy a későbbi futások folyamán ne kelljen minden értéket újból és újból kiszámolni. Ha például csak az eredmények kiszámításának módszerén szerettük volna változtatni, vagy másképpen kiszámolni az összevont értéket, akkor a meglévő terminusjelöltek további adatai már rendelkezésre álltak. Valamint ha táblázatos formában szeretnénk megnézni az eredményeket, számításokat, akkor is ez bizonyul a legjobb választásnak. Ezen kívül például a C-értékeket a mintaillesztési folyamat közben már ki tudtuk számolni, a többit pedig csak ez után, így takarékosabb volt minden számítást időben elvégezni, és az adatokat akkor elmenteni.

Az adatokat tartalmazó adatbáziskezelő-rendszer kiválasztáskor az alábbi szempontoknak kellett megfelelni: (1) ne szerveren (amely lehet akár virtuális is) futó alkalmazás legyen, hanem olyan, amelynek adatfájljait a programcsomag könyvtárába el lehessen helyezni. (2) Java nyelvre létezzen hozzá csatlakozófelület, amely, ha lehet, minél egyszerűbb legyen. (3) A programcsomag nyílt hozzáférésű legyen, ne kelljen hozzá licenszet beszerezni. (4) A programkódon belül könnyen lehessen beilleszteni UTF-8 típusú karaktereket, vagy legalább a francia nyelv speciális karaktereit (pl. *è*, *ç*, *ô*), valamint (5) a programból könnyen lehessen adatbázis-műveleteket végrehajtani, például lekérdezést, módosítást.

Ezen kritériumok mindegyikének az SQLite felelt meg, amely szerver nélküli adatbázisok tárolásához használatos adatbázismotor. Telepítése egyszerű, nyilvános forráskódú, kis méretű csatlakozófelülettel rendelkezik. Az SQLite telepítéséhez csak

egy .exe fájlra van szükség, amelynek elindításával létrehozható és módosítható egy adatbázis. Főbb hátránya a gyenge típusosság (pl. egy számot váró mezőbe be lehet írni szöveget is), de ez az adatok felvitelekor programkódos ellenőrzéssel megoldható (Allen és Owens 2010, Kreibich 2010). A létrehozott adatbázis kis méretű, tömör, ugyanúgy mint az illesztőprogram, így az új okostelefonokon használt Android operációs rendszer előnyben részesített adatbázisa lett (Murphy 2010).

Az adatbázis létrehozásakor ügyeltünk arra, hogy az mind a három normálformának megfeleljen (Garcia-Molina és mtsai 2002). Az 1. normálforma (1NF) szerint egy adatbázisban az egyes cellákban csakis egyszerű adatok lehetnek, tehát például listák, felsorolások nem. Például ha egy személynek több lakhelye is van, akkor azt mindenképpen külön adatsorban tároljuk.

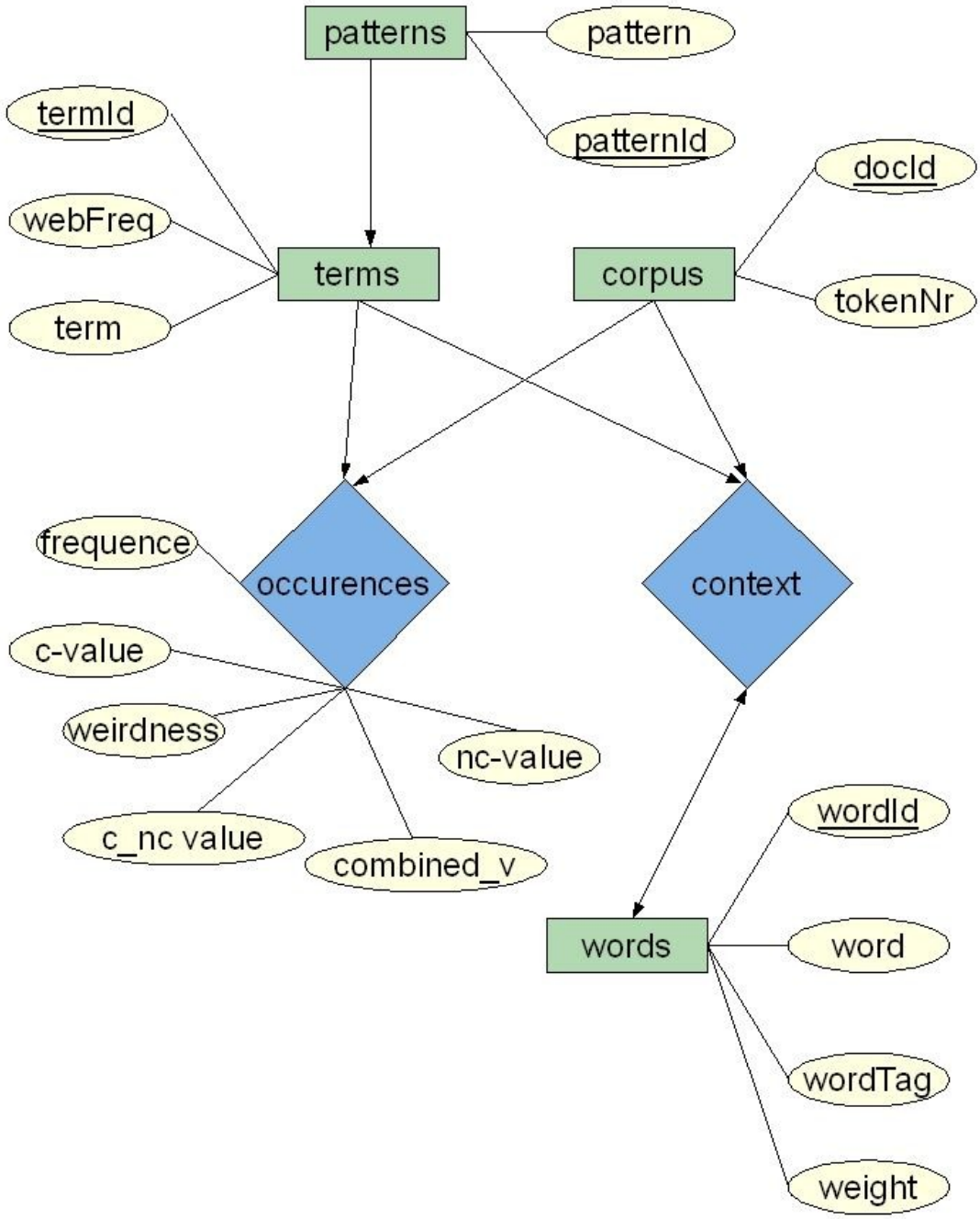
A 2. normálforma (2NF) szerint egy adatbázis akkor megfelelő, ha azon elemeket, amelyek nem kulcsok, ez utóbbiak teljesen meghatározzák. A kulcsok az adatbázis azon elemei, amelyek egy adatbázis egy egyedét egyértelműen meghatározzák, például az emberek esetében a név nem lehet kulcs, a személyi szám igen, mert több embernek is lehet ugyanaz a neve, de a személyi számuk nem egyezhet meg. Eszerint nem szabad olyan értékeket tartalmaznia egy adatsornak, amely nem köthető egyértelműen az adott egyedhez, kulcshoz. Például ne tároljunk egy adatbázisban egy személynél a személy egyedi adatait (pl. adószám, név, születési dátum) a nem egyedi adataival, mint például cím, amelyből több is lehet, mert az az egyedi adatok ismétlését vonná maga után.

A 3. normálforma (3NF) szerint az adatbázisban nem lehet tranzitív függés, azaz minden olyan elemnek, amely nem kulcs, közvetlenül kell függnie a kulcstól. Például ne tároljuk egy vállalkozó személyi adataival egy sorban a fő vállalkozási területét, és annak kódját, mert a kód egyértelműen beazonosítja a fő tevékenységi kört, így a személyi számtól függne a fő vállalkozási terület kódja, amelytől pedig a vállalkozási terület függne. Ilyen esetben ezt is két adattáblára kell bontani.

A 3NF-ben lévő adatbázisunk hat táblából áll, amelyek közül kettő kapcsolati tábla, tehát legalább két táblát köt össze. Az első tábla tartalmazza a terminusjelölteket, azok kódjával, és a köznyelvi korpuszbéli előfordulási számukkal, amely azért került ebbe a táblába, mert azt a terminusjelölt egyértelműen meghatározza. A terminusjelöltek tartalmazzák még saját morfoszintaktikai mintájuknak (pl. N N N) kódját is, amit a *patterns* tábla ír le. A *corpus* tábla a feldolgozott szabadalmak adatait tartalmazza: a szabadalom kódját (amely azt egyértelműen meghatározza), és a benne előforduló tokenek

számát, ami a gyakorisági értékek kiszámításánál fog szerepet játszani. Az *occurrences* tábla egy kapcsolati tábla: egy adott terminusjelölt adott korpuszbeli előfordulásának adatait tartalmazza: gyakoriság, C-érték, *weirdness*, C/NC-érték, összevont érték és súlyérték. Ezen mértékek mind egy adott terminusjelölt egy adott korpuszban előforduló értékeit képviselik, így azokat nem közvetlenül a terminusjelölthöz, hanem ehhez a táblához vettük fel. A kulcs nélküli *context* tábla egy adott terminusjelölt egy adott szövegbeli adott szókörnyezetét jeleníti meg. A szókörnyezetet a *words* táblában lévő elemek alkotják, amelyek tartalmazzák magát a szótövesített szóalakot, ennek szintaktikai kódját (azért, hogy például megkülönböztessük az *être* főnevet az *être* igétől), majd pedig ezek súlyát, amely az adott környezetszónak van. A *context* tábla egy sora tehát azt mutatja meg, hogy az adott dokumentumban az adott terminusjelölt előtt és után milyen szavak találhatóak minden egyes előfordulásnál.

Az előbb említett adatbázis egyed-kapcsolat diagramját a 7.6. ábra mutatja be.



7.6. ábra: A terminusjelöltek adatait tartalmazó adatbázis E-K diagramja

8. Eredmények

A 8. fejezetben először azt ismertetjük, hogyan történt az eredmények kiszámítása (8.1), majd azt, hogy a TE első fázisában, azaz a szabály alapú kinyerés és szűrés során milyen értékeket kaptunk az informatikai korpuszon (8.2.). A 8.3. alfejezetben azt mutatjuk meg, hogy a különböző statisztikai mértékek hogyan járultak hozzá a TE hatékonyságának növeléséhez, és hogy ehhez milyen küszöbértékeket kellett választani dokumentumonként, illetve az összes dokumentumra. A 8.4. alfejezetben azt ismertetjük, milyen hatékonyságúak az informatikai korpuszra alkalmazott módszerek egy másik szakterületen. A 8.5. fejezet célja a terminológiakivonatoló eredményeinek értelmezése a fontosabb hibák megemlítésével. Végül a 8.6. fejezetben történik a saját terminológiakivonatoló eredményeinek összevetése más terminológiakivonatolókkal (Fastr és YaTeA).

8.1. Az eredmények számításának módszere

A TE esetében az eredmények kiszámításakor két főbb értéket kell meghatározni: a pontosságot és a fedést. Az első azt mutatja meg, hogy a kinyert terminusok között milyen arányban szerepelnek valódi terminusok, a második azt, hogy a valódi terminusok terminusok hány százalékát nyerte ki az alkalmazás. A harmadik érték, az F-érték, az előbbi kettő harmonikus közepe, így az előbbi kettő ismeretében már megállapítható. Az eredmények kiszámításához szükség van a kivonatoló által adott összes terminusjelöltre, valamint arra a listára, amely az adott korpuszból tartalmazza az összes terminusjelöltet. Ez utóbbi listát előzetesen kézzel állítottuk össze a korábban megadott terminusdefiníciók alapján. A fedés és a pontosság kiszámításakor az alkalmazás a terminusjelöltek egyszeri előfordulását veszi alapul, azaz ha egy terminusjelöltet sikeresen felismer, akkor lényegtelen, hányszor fordul elő az adott dokumentumban, csak egynek számít.

Minthogy a korpuszban előforduló összes terminus rendelkezésre áll a terminológiakivonatoló alkalmazás számára, így mind a fedés, mind a pontosság értékeit a program automatikusan számolja ki, miután az előző fejezetben említett adatbázist minden gépi terminusjelölttel, valamint az ezekhez tartozó statisztikai értékekkel (pl. *weirdness*) már feltöltötte.

A programban a validálást elvégző osztály foglalkozik a statisztikai módszerek hatékonyságának számításával is. A statisztikai módszerek ismérve, hogy minden egyes terminusjelölthöz egy értéket rendelnek. Feltételezhető, hogy minél nagyobb az adott

terminusjelölthöz hozzárendelt bármelyik korábban ismertetett statisztikai érték (C-érték, *weirdness*, súly, valamint az ezekből képzett összevont érték), annál nagyobb valószínűséggel terminus az adott jelölt. Ezért a validálás során meg kell határozni egy küszöbértéket a statisztikai értékeknél, amely felett a legnagyobb a fedés, a pontosság vagy az F-érték. A küszöbérték meghatározása is automatikusan történik a terminusjelöltek statisztikai értékei alapján.

8.2. Szabály alapú kinyerés és szűrés eredményei

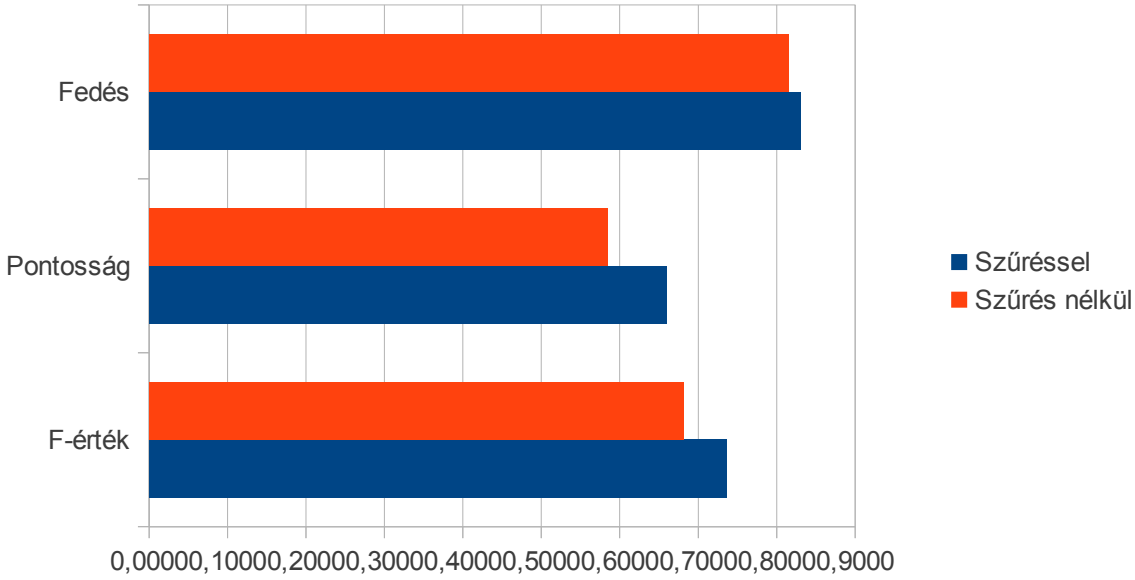
A 8.1. táblázatban bemutatjuk az informatikai korpuszból történő szabály alapú terminuskinyerés és a szabály alapú szűréssel kibővített terminuskinyerés eredményeit. Mivel a szabály alapú szűrés a szabály alapú kinyerés előtt hajtódott végre, azaz nem a már kinyert terminusokat szűrtük tovább, így a szűréssel akár a fedés is nőhet. A táblázat soraiban a dokumentumok nevei szerepelnek, és az abban elért fedés, pontosság és F-érték eredményei. A legutolsó sor a 10 dokumentumra kivetett átlagos értéket mutatja.

8.1. táblázat: A különböző dokumentumokban elért fedés, pontosság és F-értékek szabály alapú szűrés nélkül, majd szűréssel

	Szűrés nélkül			Szűréssel		
Dokumentum	Fedés	Pontosság	F-érték	Fedés	Pontosság	F-érték
FR2008051044	0,7692	0,3937	0,5208	0,7923	0,5282	0,6338
FR2008051104	0,8135	0,5358	0,6461	0,8497	0,6332	0,7257
FR2008051812	0,7789	0,5441	0,6407	0,8105	0,6638	0,7299
FR2008051823	0,8356	0,5914	0,6926	0,8691	0,6745	0,7595
FR2008051836	0,8485	0,6863	0,7588	0,8818	0,8039	0,8410
FR2008051856	0,8217	0,4907	0,6145	0,8682	0,6087	0,7157
FR2008051890	0,7672	0,5017	0,6067	0,7937	0,5976	0,6818
FR2008052025	0,8056	0,5377	0,6450	0,8317	0,6409	0,7240
FR2008052073	0,7710	0,4755	0,5882	0,7897	0,5808	0,6693
FR2008052077	0,8161	0,5657	0,6682	0,7931	0,6330	0,7041
Átlag	0,8027	0,5323	0,6382	0,8280	0,6365	0,7185

A táblázatból jól látható, hogy egy ilyen erősen szakmai korpuszban már a szabály alapú kinyerés is eredményes lehet, hiszen az abban szereplő főnévi elemek nagyobbrészt terminusok, illetve a főnévi terminusok többsége illeszkedik az előre megadott mintákra. Azonban a szűrés nélküli algoritmus pontossága (0,53) nem megfelelő, ugyanis ez az érték azt mutatja, hogy annak ellenére, hogy a szaknyelvben nagy arányú a terminusok száma, a terminusjelöltek mégis elég sok nemterminusi főnevet, illetve olyan melléknevet tartalmaznak, amelyek nem részei a terminusoknak.

Mint ahogy a 7.2.2. fejezetben is említettük, a szabály alapú szűrés esetén kiküszöbölhetjük azon főneveket, mellékneveket, amelyek nem lehetnek terminusok részei. Ilyenek voltak például a *par exemple*, *en effet* konnektívumok vagy a *suivant* melléknév. A fenti táblázatból és a 8.1. ábrából jól látszik, hogy a szabály alapú szűréssel a fedést már nagyon nem tudtuk növelni, ellenben a pontosság értékei 10%-kal, az összevont mutató, az F-érték 8%-kal emelkedett. Így hipotézisünk, miszerint a szabály alapú szűréssel a pontosság jelentősen növelhető, beigazolódott.



8.1. ábra: Fedés, pontosság és F-értékek alakulása szabály alapú szűréssel, illetve anélkül

8.3. A terminológiakivonatolás hatékonyságának növelése statisztikai módszerekkel

A 7.5. fejezetben említett statisztikai módszerek közül a súly-, a *weirdness*- és a C-értékeket vettük figyelembe, egészen pontosan ezek összevont értékét. Ehhez először normalizáltuk az egyes terminusjelöltekhez tartozó ezen értékeket, leképeztük őket a [0..1] intervallumba a 7. alfejezet által leírtaknak megfelelően. Ezt követően létrehoztunk belőlük egy összevont értéket, majd megnéztük, hogy ezek esetében mekkora küszöb megválasztásánál volt a legmagasabb az F-érték. A maximális F-érték és pontosság eléréséhez a megválasztott küszöbérték dokumentumonként eltér. A terminológiakivonatoló élesben történő használatához ezekből egy értéket kell képezni, és azt alkalmazni a dokumentumok összességére.

8.3.1. Az egyes értékek súlyozása az összevont érték esetében

A statisztikai mértékek jellegzetessége, hogy általában a pontosságot növelik a fedés kárára. Mivel a szabály alapú terminuskinyerés esetében magas fedés és alacsony pontosság várható, a statisztikai mértékeknél pedig alacsony fedés magas pontossággal, így a három statisztikai értékekből egy olyan összevont értéket kell létrehozni, amely a szabály alapú terminuskinyerés alacsonyabb pontosságát jelentősen növeli a fedés lehető legkisebb csökkentésével. Mivel a fedés és a pontosság együttes mutatója az F-érték, így a cél az F-érték növelése.

A statisztikai módszerek súlyozásához az első öt, informatikai szövegen végeztünk kísérletet. A három statisztikai mértéket szinte az összes lehetséges kombinációban kipróbáltuk, és megmértük, hogy az egyes kombinációkban mennyi lett az öt dokumentumban a terminusok ezen összevont értéke, majd megmértük, hogy mely esetekben lett maximális az összes dokumentum F-értéke. Ehhez külön kellett venni az egy- és a többszavas jelölteket, amelyek közül az utóbbiakra a C-érték is alkalmazható, míg az előbbire csak a *weirdness*- és a súlyérték. A 8.2. táblázat foglalja össze a legjobb F-értéket elérő kombinációkat az egy- illetve a többszavas terminusok esetében.

8.2. táblázat: Az egy-, illetve többszavas terminusok kinyeréséhez használt statisztikai mértékekkel elérhető legnagyobb F-értékek

Egyszavas terminusok		Többszavas terminusok			F-érték
Súly	Weirdness	Súly	Weirdness	C-érték	
0,2	0,8	0,8	0,2	0	0,7275
0,2	0,8	0,6	0,4	0	0,7272
0,2	0,8	0,7	0,3	0	0,7272
0,1	0,9	0,8	0,2	0	0,72719
0,1	0,9	0,9	0,1	0	0,7270

A fenti táblázatból jól látható, hogy az egyszavas terminusok esetében a weirdness-érték a legmeghatározóbb: ha ezeket 0,8 és 0,9-re választjuk, akkor érhető el a legnagyobb F-érték. A többszavas terminusok esetén viszont éppen a súly értékei meghatározóbbak: minél nagyobbnak választjuk ezt meg, az F-érték annál inkább nő. A C-értékekkel kapcsolatban meglepőek az eredmények: legtöbbször nem járul hozzá az F-érték növeléséhez.

Ezek az eredmények igazolják azon hipotézisünket, hogy a *weirdness*-érték elsősorban az egyszavas terminusok esetén lehet hatékony, ugyanis az általunk használt

mérési módszerrel a többszavas terminusokra nehezebb megállapítani ezt az értéket, mert ezekre az internetes keresőmotorokkal nem lehet pontos értéket adni. Azt ugyanis meg lehet adni a kereséskor, hogy az elemek egymáshoz közel álljanak, de azt nem, hogy hány elem állhat közöttük, vagy hogy közvetlenül egymás mellett kell-e lenniük, így az adott többszavas terminusra történő keresés nem csak azon oldalak számát adja vissza, amelyekben pontosan azok a többszavas terminusok szerepelnek.

A későbbiekben a legnagyobb F-értéket elért kombinációt használjuk fel az összevont érték kiszámításához, azaz egyszavas terminusoknál a súly 0,2, a *weirdness* 0,8, többszavas terminusoknál a súly 0,8, a *weirdness* 0,2, a C-érték pedig 0 arányban járul hozzá ezen érték kiszámításához.

8.3.2. Statisztikai értékek hatása a terminológiakivonatolás eredményeire

A statisztikai értékektől azt várjuk el, hogy a fedés értékét nem jelentősen csökkentve a pontosságot, és ebből adódóan az F-értéket is növelik. Először a pontosság javításában elért eredményeket mutatjuk be a 8.3. táblázatban.

8.3. táblázat: Legnagyobb pontosság (illetve az ahhoz tartozó fedés és F-érték) értékek a különböző dokumentumokban

Dokumentum	Fedés	Pontosság	F-érték	Küszöb
FR2008051044	0,0154	0,6667	0,0301	0,8982
FR2008051104	0,0466	0,9	0,0887	0,8859
FR2008051812	0,0316	1,0000	0,0612	0,8261
FR2008051823	0,0168	1,0000	0,0330	0,8261
FR2008051836	0,0273	1,0000	0,0531	0,8843
FR2008051856	0,0078	1,0000	0,0154	0,8261
FR2008051890	0,0212	0,8	0,0412	0,8286
FR2008052025	0,0189	0,8605	0,0370	0,8982
FR2008052073	0,0187	0,6667	0,0364	0,8886
FR2008052077	0,0168	1,0000	0,0330	0,8261
Átlag	0,0221	0,8992	0,0429	0,8588

A táblázat értékeiből jól látszik, hogy a statisztikai értékek a pontosságot nagymértékben tudják növelni (akár 1 is lehet ez az érték), de ezt csak a fedés és az F-érték kárára, melyek ezáltal nagyon alacsonyak lesznek. A maximális pontosság elérése érdekében a küszöböt elég magasra kell állítani, így viszont sok terminust is kiszűrünk, így mind az F-érték, mind a fedés jelentősen csökken.

Második lépésben megvizsgáltuk, hogy milyen mértékű F-érték javulás érhető el a statisztikai módszerek segítségével. Ezt mutatja a 8.4. táblázat a legnagyobb elért F-értékekkel.

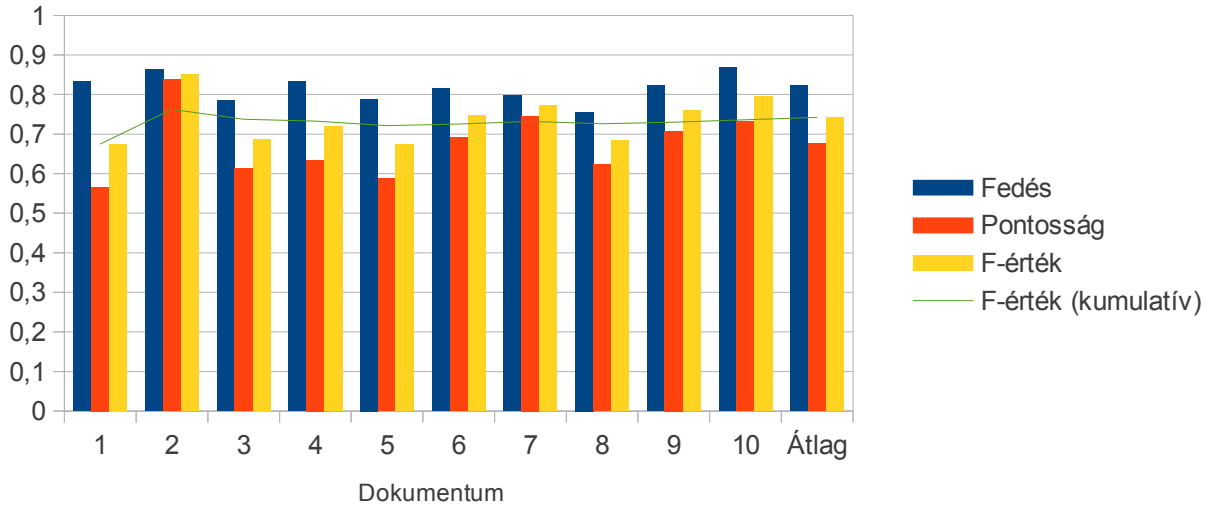
8.4. táblázat: Legjobb F-értékek a statisztikai módszerekkel

Dokumentum	Fedés	Pontosság	F-érték	Küszöb
FR2008051044	0,7923	0,5450	0,6458	0,0297
FR2008051104	0,8342	0,6518	0,7318	0,0488
FR2008051812	0,8105	0,6937	0,7476	0,0443
FR2008051823	0,8591	0,6938	0,7676	0,0216
FR2008051836	0,8788	0,8101	0,8430	0,0178
FR2008051856	0,8682	0,6588	0,7492	0,0471
FR2008051890	0,7937	0,6148	0,6928	0,0139
FR2008052025	0,8251	0,6565	0,7312	0,0235
FR2008052073	0,7804	0,6029	0,6802	0,0520
FR2008052077	0,7931	0,6449	0,7113	0,0133
Átlag	0,8235	0,6572	0,7301	0,0312

A 8.4. táblázat eredményeiből jól látszik, hogy a 10 dokumentumot körbefogó *Átlag* értéknél nincs jelentősebb eltérés a szabály alapú szűrés által létrehozott eredmények, illetve a statisztikai módszerek eredményei között, a javulás mindössze 1,2%-os (0,7185 → 0,7301).

8.3.3. A korpusz nagyságának megfelelősége

A korpusz méretének összeállításakor fontos volt meghatározni, hogy az egyes területekről hány darab szabadalmi leírást kellett választani ahhoz, hogy a terminológiakivonatoló eredményeinek vizsgálatakor reprezentatív értéket kapjunk. Ehhez megnéztük a 10 informatikai témájú szabadalmi leíráson a statisztikai módszerek eredményességét az F-érték szempontjából. Azaz megnéztük, hogy a statisztikai mértékek alkalmazásával kapott eredmények hogyan alakulnak az egyes dokumentumok után. A fő értéknek azért az F-értéket választottuk, mert ez az az érték, amelyik mind a fedés, mind a pontosság értéket magában foglalja. Ezt a 8.2. ábra szemlélteti, ahol a vízszintes tengelyen a dokumentum száma szerepel, a függőleges tengely mutatja a fedés, pontosság és F-értékeket, a vonal pedig a kumulatív F-értéket, ami minden egyes dokumentumnál annak és az összes többi előtte lévő dokumentum F-értékének átlaga.



8.2. ábra: Pontosság, fedés és F-értékek alakulása az egyes dokumentumok feldolgozása után

Az ábrából jól látható, hogy a 4. dokumentumtól számítva a 0,73 körüli érték állandónak tekinthető, így több dokumentumból álló korpusz választása esetén nem valószínűsíthető az érték jelentős javulása vagy romlása.

8.4. A terminológiakivonatolás eredményei az A23L korpuszon

A terminológiakivonatoló alkalmazást az A23L szabadalmi osztályból választott szövegeken is lefuttattuk. Ennek célja az volt, hogy megmutassuk, van-e hatékonyságbeli különbség a két eltérő szakterületből választott korpusz között.

A 8.5. táblázatban bemutatjuk az A23L korpuszból történő szabály alapú terminuskinyerés és az azt követő szűrés eredményeit. A táblázat soraiban a dokumentumok nevei szerepelnek, és az abban elért fedés, pontosság és F-érték eredmények. A legutolsó sor a 10 dokumentumra kivetett átlagos értéket mutatja.

8.5. táblázat: Az A23L dokumentumokban elért fedés, pontosság és F-értékek szabály alapú szűrés nélkül, majd szűréssel

	Szűrés nélkül			Szűréssel		
Dokumentum	Fedés	Pontosság	F-érték	Fedés	Pontosság	F-érték
EP2008056887	0,7609	0,4930	0,5983	0,7826	0,5806	0,6667
EP2008061497	0,7007	0,4392	0,5399	0,7270	0,4977	0,5909
EP2008063597	0,7636	0,6400	0,6963	0,7852	0,6935	0,7365
EP2009057808	0,7593	0,5256	0,6212	0,7963	0,6099	0,6908
FR2006001856	0,7679	0,5482	0,6397	0,7848	0,6327	0,7006
FR2007001526	0,7462	0,6599	0,7004	0,7308	0,6835	0,7063
FR2007051158	0,6849	0,5263	0,5952	0,7466	0,6301	0,6834
FR2007051178	0,7443	0,5377	0,6243	0,7705	0,6138	0,6833
FR2007051372	0,7137	0,5156	0,5987	0,7608	0,6178	0,6819
FR2007051549	0,6961	0,5272	0,6	0,7293	0,6168	0,6684
Átlag	0,7337	0,5413	0,6238	0,7614	0,6176	0,6809

A táblázat alapján a csupán szabály alapú terminuskinyerés az A23L korpuszon is elég nagy fedést (0,7337) biztosít alacsony pontossággal (0,5413). Az első érték kissé alacsonyabb az informatikai korpusz esetén mért értéknél (0,8027), de a pontosság nagyjából hasonló mindkét esetben. Ez azt mutatja, hogy már a szabály alapú terminuskinyerés is nagy fedést biztosít ezen a területen is.

A szabály alapú szűrés itt is hasonló eredményt hozott, mint az informatikai témájú korpuszon. A fedést a szabály alapú szűrés csak minimális mértékben növelte (kb. 0,03-dal), a pontosságot már nagyobb mértékben: a szabály alapú kinyerés 0,5413 értékét 0,6176-re növelte, azaz körülbelül 0,07-dal. Az F-érték a pontosság és a fedés növekedésének függvényében kb. 0,06-dal nőtt.

A statisztikai módszerek esetében az A23L korpuszon is az feltételezhető, hogy jelentősen csak a pontosságot tudják növelni, a fedést nem. A 8.6. táblázatban mutatjuk be az egyes dokumentumokon elért (és az átlag) legjobb pontosság értékeket. A küszöbérték azt mutatja meg, hogy a kombinált érték esetében mely volt az az érték, amely felett azt az eredményt kaptuk.

8.6. táblázat: Legnagyobb pontosság (illetve az ahhoz tartozó fedés és F-érték) értékek a különböző dokumentumokban

Dokumentum	Fedés	Pontosság	F-érték	Limit
EP2008056887	0,0326	0,8571	0,0628	0,9026
EP2008061497	0,0493	0,6522	0,0917	0,8261
EP2008063597	0,0846	0,9512	0,1554	0,8886
EP2009057808	0,0231	1,0000	0,0452	0,9005
FR2006001856	0,0886	1,0000	0,1628	0,8261
FR2007001526	0,0692	1,0000	0,1295	0,8261
FR2007051158	0,0685	0,9091	0,1274	0,9005
FR2007051178	0,0492	0,8369	0,0930	0,9026
FR2007051372	0,1451	0,6379	0,2364	0,6770
FR2007051549	0,0497	0,9	0,0942	0,8977
Átlag	0,0660	0,8716	0,1199	0,8548

A táblázat értékeiből jól látható, hogy a statisztikai értékekkel a megfelelő küszöb meghatározásával a pontosság valóban jelentős mértékben növelhető: 3 dokumentumban is sikerült elérni a 100%-os arányt – természetesen ekkor a fedés jelentősen csökkent. Az átlagos pontosság értéke a 10 dokumentumra levetítve 0,8716: a fedés átlaga ekkor 0,0660. Második lépésben megvizsgáltuk, hogy milyen mértékű F-érték javulás érhető el a statisztikai módszerek segítségével. Ezt mutatja a 8.7. táblázat.

8.7. táblázat: Legnagyobb F-érték (illetve az ahhoz tartozó fedés és pontosság) értékek a különböző dokumentumokban

Dokumentum	Fedés	Pontosság	F-érték	Küszöb
EP2008056887	0,7228	0,6364	0,6768	0,1054
EP2008061497	0,7237	0,5	0,5914	0,0087
EP2008063597	0,7831	0,7023	0,7405	0,0058
EP2009057808	0,7963	0,6370	0,7078	0,0430
FR2006001856	0,7637	0,6582	0,7070	0,1097
FR2007001526	0,7308	0,6985	0,7143	0,0115
FR2007051158	0,7466	0,6450	0,6921	0,0111
FR2007051178	0,7651	0,6219	0,6861	0,0124
FR2007051372	0,7608	0,6198	0,6831	0,0061
FR2007051549	0,7182	0,6311	0,6718	0,0605
Átlag	0,7511	0,6500	0,6969	0,0374

A 8.7. táblázat jól mutatja, hogy az A23L korpuszon sem tudták a statisztikai mértékek jelentősen megnövelni az F-értékeket. Az átlagos F-érték 0,6969 csak kb. 0,02-del nagyobb a szabály alapú szűrés utáni F-értéktől.

8.5. A terminológiakivonatolás eredményeinek elemzése

A 8. fejezet célja, hogy megmutassa, milyen hatékonyságot ért el az általunk kidolgozott terminológiakivonatoló. Azonban a fedés, pontosság és F-értékek felsorolása a különböző esetekben nem lehet elegendő: azt is megvizsgáljuk, hogy az egyes módszereknél melyek voltak azok a terminusok, amelyek a szűrés ellenére mégsem kerültek be a terminusjelölt-listába, és melyek azok a nem terminusok, amelyek a különböző módszerek ellenére mégis bennmaradtak a listában.

Ehhez mindkét szaknyelv (informatikai és alapvető emberi szükségletek) korpuszából választunk egy-egy szabadalmi leírást, amelyben az összes hibás esetet végignéztük. A G06F-korpuszból az FR2008051823 számú dokumentumot, az A23L-korpuszból az FR2007051158 számú dokumentumot választottuk. Az első dokumentum 2760 szövegtokenből áll, és 193 kézzel annotált terminussal rendelkezik. A második dokumentum mérete 4372 szövegtoken, és a kézi annotációk alapján 462 különböző terminussal rendelkezik.

8.5.1. A szabály alapú kinyerés és szűrés hatékonysága

A korpuszt bemutató 6. fejezetben részletesen ismertettük, miért választottunk korpusznak szabadalmi leírásokat: (1) terminusok pontos használata, és ezáltal (2) sok ismétlődés, amely a gyakoriság alapú módszerek számára igen kedvező. A korpusz kézi feldolgozásakor (a terminusok kézi bejelölésénél) már észrevehető volt, hogy a szabadalmi leírások ritkán tartalmaznak nem terminusi főnévi szerkezeteket, így ezen szövegekre már a szabály alapú módszerek is igen hatékonyan működnek: ami a megadott főnévicsoport-mintáknak megfelel, az valószínűleg terminus. Sok esetben a leírások szinte csak terminusokat tartalmaznak, erre példa az egyik szabadalomból idézett mondat:

> (1) Un système d'annotation est un système qui permet d'ajouter de l'information de haut niveau appelée métadonnées sur un document multimédia, c'est-à-dire un document textuel, d'image, audio et/ou vidéo.[29]

Nagyon ritka volt azon mondatok száma, amelyek nagyobbrészt nem terminust tartalmaztak, erre példa lehet (2):

> (2) Une augmentation pondérale entraîne en effet, en plus des désagréments moraux (problèmes psychologiques) liés aux dictats de la minceur et de la beauté des sociétés modernes, des problèmes de santé.[30]

[29]Az annotációs rendszer egy olyan rendszer, amely lehetővé teszi a metaadatnak nevezett magas szintű adatok hozzáadását multimédia-dokumentumokhoz, azaz szöveges, kép-, hang- és/vagy videódokumentumokhoz. (saját ford.)

A (2)-vel jelölt mondat az *augmentation pondérale* 'testsúlynövekedés', *problème psychologique* 'pszichés probléma' és a *problème de santé* 'egészségügyi probléma' kivételével nem tartalmazott terminust. A többi főnévi elem ugyanis nem köthető a szaknyelvhez: *minceur* 'karcsúság', *beauté* 'szépség' vagy *société moderne* 'modern társadalom'.

A G06F-dokumentum 133 valós terminust tartalmaz, amelyből a szabály alapú kinyerés 111-et talált meg (így a fedés 0,834). A pontosság tekintetében azonban már nem olyan jók az eredmények: a szabály alapú kinyerés alapján ebben a dokumentumban 216 terminus szerepel, így 115 felesleges elemet tartalmaz (a pontosság tehát 0,5139). Az előbbi két értékből számolt összevont mutató, az F-érték így 0,6361.

A G06F-dokumentum szabály alapú szűrésével mind a pontosság, mind a fedés nőtt: az előbbi jelentős mértékben, az utóbbi kevésbé. A helyesen kinyert terminusok száma 111-ről 134-re nőtt, a fedés így 0,857. A terminusjelölt-listát 216-ról sikerült 190-re csökkenteni, így a pontosság 0,6-ra nőtt. Az F-érték kb. 7%-kal nőtt, mert ekkor az értéke 0,7059.

Az A23L dokumentumában a szabály alapú kinyerés a 461 terminusból 352 terminust nyert ki helyesen, így a fedés 0,7636. Ezen szakaszban a terminológiakivonatoló szerint 550 terminus van a dokumentumban, így 198 nem terminust is annak vélt, ezért a pontosság itt 0,64. A fedés és pontosság harmonikus közepe, az F-érték a kinyerés esetében 0,6963.

A szabály alapú szűrés itt is növelte mind a pontosság, mind a fedés értékeket, amelyek közül az utóbbit nagyobb mértékben. A szabály alapú szűréssel a helyesen kinyert terminusok száma 362 (a korábbi 352 helyett), így a fedés már 0,7852. A terminusjelöltek száma 550-ről 526-ra csökkent a helyesen kinyert terminusok növekedésével, így a pontosság értéke is nőtt, az új pontosság 0,688. Az F-érték kb. 4%-kal emelkedett, mivel az új F-érték 0,7335.

A szabály alapú kinyerés eredménye valóban igazolta azon hipotézisünket, hogy már önmagában a szabály alapú kinyerés nagy fedést eredményez. A pontosság kb. 0,5 és 0,6 közé eső értéke azonban azt mutatja, hogy mégis sok olyan főnévi vagy melléknévi bővítményelemet tartalmaz egy szabadalmi leírás, amely nem lehet terminus.

[30]A testsúlynövekedés ugyanis nemcsak a modern társadalmak szépséggel és karcsúsággal kapcsolatos elvárásainak betudható morális kellemetlenségeket (pszichológiai problémákat) von maga után, hanem egészségügyi problémákat is. (saját ford.)

8.5.2. A szabály alapú kinyerés és szűrés lehetséges hibaforrásai

Mint ahogy az eredményekből is látható, a szabály alapú kinyerés és szűrés önmagában nem lehet elegendő még egy ennyire szakmai jellegű korpuszban sem. A továbbiakban azt mutatjuk be, melyek azok a hibaforrások, amelyeket a szabály alapú módszerek hordoznak magukban.

8.5.2.1. Helytelen morfológiai annotációk

Mivel a terminológiakivonatoló elsősorban szabály alapú módszerekkel történik, azaz adott morfoszintaktikai mintákkal, ezért fontos, hogy az automatikusan működő POS-tagger minél hatékonyabban működjön. Azonban minden annotáló program egy kisebb-nagyobb hibaszázalékkal dolgozik, ezért az ebből fakadó hibák kikerülhetetlenek.

Az informatikai korpuszban a fel nem ismert terminusok közül (19) 5 esetben volt ez a hiba oka, a terminusjelölt-listába 6 került feleslegesen emiatt (76-ból). Az A23L korpuszon az arány másképp alakult, mert itt 31 esetben azért nem ismerte fel a terminust, mert rossz volt a *POS-tag*e (a fel nem ismert terminusok száma 99), és 17-szer került felesleges elem a listába (164 eset közül).

Gyakori eset a főnevek mellékneveknek jelölése, például a *solvant* ’oldat’ helyett ’oldó’ melléknévként lett jelölve, hasonlóan a *terminal* szóhoz, amely lehet ’terminál’ főnév vagy ’végső’ melléknév is, a szövegben viszont mindkettő csak főnévként volt használva. Ezen kívül ugyanez fordítva is gyakran megesik, például az *anti-oxydant* ’antioxidáns’ a szövegben főnévként szerepelt, de az annotáció szerint melléknév, hasonlóan a *minéral* szóhoz ’ásvány’ vagy ’ásványi’ jelentésben.

A 2. melléklet táblázatában külön jelöltük azokat az eseteket, amelyeket nem tekinthetünk mindig egyértelmű hibának. Sok melléknévi igenévi alak használatos melléknévként is, így ezen esetekben a tévedés nem mindig tudható be egyértelműen a *POS-tagger*nek. Ezek megkülönböztetése ugyanis sokszor a kézi annotálók számára sem egyértelmű. A fenti két korpuszban például a *comprimé enrobé* ’bevont tabletta’ terminus, de a *médicament usuellement utilisé* ’rendszeresen használt gyógyszer’ nem, mert az utóbbi esetben nem melléknév, hanem melléknévi igenév. Egy-két esetben azonban a kézi annotálás számára egyértelmű igeneveket is melléknévként jelölt be az automatikus annotáló, például a *paiement électronique comprenant* ’...-t tartalmazó elektronikus fizetés’ vagy a *polysaccharide contenant* ’...-t tartalmazó poliszacharid’, amelyek esetében az *-ant* végződés fejezi ki a folyamatos melléknévi igenevet.

A hibás morfológiai címkézés másik esete, amikor nem a szófaji címke, hanem a szótövesített alak a nem megfelelő. Azaz a terminusjelölt lehet, hogy jó, de rossz a szótöve, így nem egyezik a kézzel annotált listában szereplővel. Erre csak egy esetet találtunk: a *POS-tagger* az informatikai korpuszban a *fils* 'vezetékek' szót a *fils* 'vkinek a fia' szótővel társította, nem pedig a *fil* 'vezeték' szótőhöz.

8.5.2.2. Egyéb hibaforrások

Természetesen nem minden terminus, ami a megadott szintaktikai mintára illeszkedik: például az egyszavas főnevek gyakran lehetnek köznyelvi egységek is. Az informatikai példaszöveg esetében 39 olyan terminusjelölt szerepel, ami valójában köznévi egység, az A23L szövegében pedig 52. Ilyenek a *place* 'hely', *an* 'év' stb. Ezek későbbi szűrésére használandóak a statisztikai módszerek, amelyek a szövegbeli előfordulási arányuk, környezetük alapján próbálják eldönteni, hogy az adott jelölt tényleg terminus-e.

A felsoroltakon kívül kisebb mértékben más hibaforrások is előfordulhatnak. Adódott még olyan eset is, hogy nem volt olyan minta, amely illeszkedett volna a terminusra. Például számneveket nem engedtünk meg terminusokban, mert a szó után álló szám legtöbbször ábrahivatkozást jelöl. A *terminal 2* 'a 2 jelű terminál' vagy a *câble USB 14* '14-es jelű USB-kábel' esetében a számok a mellékelt ábrára tesznek hivatkozást. Egy-két esetben a számnevek azonban a terminusok részei, például a *diabète de type 2* '2-es típusú diabétesz' vagy *oméga 3* 'omega 3 [zsírsav]' esetében.

A korpuszban ritkán fordult elő olyan eset, amikor a terminusban determináns is előfordult, így ezen hipotézisünk is igazolódott. Ezen esetek közül való a szabadalmak szövegében gyakran előforduló *homme du métier* vagy *homme de l'art* 'szakmabeli', vagy a *sensation de la faim* 'éhségérzet' vagy az *état des* (de+les) *lieux* 'tárgymutató'.

Annak ellenére, hogy a főbb típusú koordinációkat szabály alapú újraírószabályok segítségével kezeltük, előfordult, hogy a koordináció bizonyos altípusaiból származó mellérendelő szerkezeteket a program nem ismerte fel. A *document audio et/ou vidéo* 'audió- és videódokumentum' típusú mellérendeléseket a program helyesen két külön terminusként ismerte fel (*document vidéo* és *document audio*). A vesszőt is tartalmazó felsorolásokra azonban nem írtunk szabályt, mert feltételeztük, hogy az így jól kinyert terminusokat nem tudta volna ellensúlyozni az így többletként kinyert nem terminusi elemek jelenléte. A vesszőt ugyanis inkább határolóelemként célszerű használni, azaz olyan egységként, amely terminusokat választ el egymástól. Így az *arôme de café, de*

citron, de pomme, de chocolat, de vanille, de fraise 'kávé-, citrom-, alma-, csokoládé-, vaníliaaroma' esetében csak az *arôme de café* szerkezetet nyerte ki a program.

A 2. melléklet táblázata tartalmazza a további hibaforrásokat a G06F szabadalmi területről vett mintaként szolgáló szabadalmi leírás (FR2008/051836) alapján. A táblázat első oszlopa tartalmazza a terminus(jelölt) sorszámát, a második oszlop a terminus(jelölte)t, a harmadik oszlop mutatja meg, hogy az adott elem szerepel-e a tanulókorpuszban, tehát ténylegesen terminus-e. A negyedik oszlopban jelöltük, hogy a tisztán szabály alapú terminuskinyerés esetén szerepelt-e a terminusjelölt-listában, a következő oszlop azt mutatja, hogy a szabály alapú szűrés utáni listában szerepel-e. Az ezt követő oszlop írja le, hogy a nem terminus terminusjelölt milyen oknál fogva maradt a listában a szabály alapú szűrés ellenére, illetve, hogy a terminust miért nem ismerte fel a terminológiakivonatoló, ha az valóban terminus.

A 2. melléklet nyolcadik oszlopa azt mutatja meg, mely terminusjelöltek maradtak a listában, ha az összevont érték küszöbértéke úgy lett beállítva, hogy az F-érték a lehető legmagasabb legyen. A következő oszlopban azon terminusjelöltek lettek + jellel jelölve, amelyek a nagy küszöbérték esetén maradtak meg, azaz ha a legnagyobb pontosság elérését szerettük volna. A tizedik oszlop mutatja meg az adott terminusjelölt előfordulásának számát az adott szabadalmi leírásban. A többi oszlopban az adott jelölthöz tartozó statisztikai értékek olvashatók le.

8.5.3. A statisztikai módszerek eredményei

A hipotézisünk szerint a terminológiakivonatolásban a statisztikai módszerek célja a magas pontosság és az alacsony fedés elérése. Ezért, ha azokat kombináljuk a szabály alapú módszerekkel, akkor várhatóan a szabály alapú módszerek magas fedését ugyan nem, de alacsonyabb pontosságát (és ezáltal az F-értéket) növelni tudja. Az eredmények részben igazolták ezen hipotézisünket. Magas határértékek kiválasztásával a pontosságot jelentősen meg tudtuk növelni, de ekkor a fedés jelentősen csökkent. Ha pedig alacsonynak választottuk ezt a határértéket, akkor az F-értéket (és a pontosságot is) csak kisebb mértékben tudták növelni.

A G06F (informatikai) szövegben, ha az összevont értéket elég alacsonynak (0,04823), választottuk, akkor a fedés értékének megtartásával a pontosságot sikerült a szabály alapú módszerhez képest 4%-kal növelni (0,5315-ről 0,5738-re). Ez azt jelenti, hogy az alacsony összevont értékkel rendelkező elemek eltávolításával a pontosság nőtt,

tehát valóban nem terminusi elemeket távolítottunk el. Ezen nem terminusi elemek száma 15, és többek között olyan elemeket értünk rajta, mint *instant* 'pillanat', *lieu* és *place* 'hely', *arrêt* 'megállás'.

Ugyanezen a szövegen, ha az összevont értéket magasnak választottuk meg, akkor a pontosságot jelentősen meg tudtuk növelni, egészen 1-re. Ekkor viszont már csak egy terminus maradt a terminuslistában, amelynek összevont értéke 0,9, és amely tényleg terminus (*interrupteur* 'megszakító').

Az A23L szabadalmi osztályhoz tartozó szövegen, ha az összevont érték küszöbét alacsony értékre (0,0057) állítottuk, az F-érték kismértékben növekedett: kb. 0,4%-kal: így 8 nem terminust sikerült kiszűrni, mint *c* 'c', a múlt idejű melléknévi igenév helyett főnévként taggelt *reçu* 'bizonylat' vagy 'kapott' vagy a *semaine* 'hét'. Ha az összevont értéket elég magasra állítottuk (0,8898), akkor a pontosság 0,95-re növekedett. Ekkor 40 terminusjelölt maradt a listában, amelyek kettő kivétellel nem terminusok: ezek a nem létező és rosszul taggelt *mélisser* (a valóban terminus *mélisse* 'mézfű' helyett) és a *risque* 'kockázat'.

A hibák forrása ebben az esetben azért van, mert egyrészről igaz az, hogy ami kisebb arányban fordul elő egy köznyelvi korpuszban, az valóban nagyobb eséllyel terminus (ilyen például a *fil conducteur* 'vezető szál' vagy *interrupteur* 'megszakító'), azonban vannak olyan terminusok, amelyek ugyan nem terminusi jelentésben, de gyakran előfordulnak egy köznyelvi korpuszban is: ilyen a *requête* 'kérés vagy lekérdezés' vagy a *source* 'forrás'.

A szövegkörnyezet valószínűségi értékének alapján (azaz a súlyérték) nem mindig eldönthető, hogy az adott szó terminus-e vagy sem. Az aposztróf ugyan (viszonylag) egyértelműen be tudja azonosítani a terminust, mint például a *« matching »* esetében, de a gyakori terminushatárolók, például a vessző, a determináns gyakran szerepelhet nem terminusi elem közelében is. Ennek tudható be az, hogy a nem terminus *type* és *variation* mind nagy értékeket kaptak (1,81), ugyanúgy, mint a valóban terminus *formule* 'formula' és a *granule* 'granulátum'.

8.6. Összevetés más terminológiakivonatoló alkalmazásokkal

A saját terminológiakivonatoló eredményeit összevetettük más, szabadon felhasználható alkalmazásokkal. Az egyik ilyen a Jacquemin (2001) által kifejlesztett Fastr, a másik az Aubin és Hamon (2006) által létrehozott YaTeA (Yet Another Term ExtrActor). A két

program Linux alatt futtatható verzióval rendelkezik, és igényel tokenizálót, szófaji egyértelműsítőt és lemmatizálót. Mindkét program a *TreeTagger*[31] szótövesítő kimeneti formátumát fogadja el, így a két program működéséhez ezen alkalmazás francia nyelvre alkalmazható modulját telepítettük, és lefuttattuk a szabadalmi leírásokra. Ezen kimeneti fájlok szolgálnak az alkalmazások bemenetéül.

A Fastr alkalmazás célja elsősorban dokumentumok indexelése, és az azokon belüli terminusvariánsok kezelése. A szövegben nemcsak terminusokat keres, hanem felfedi azt is, hogy azok közül hány terminus szerepel a szövegben valamilyen más formában is. Ehhez az angol és a francia nyelvben részletesen megfigyelt terminusvariációk alapján adtak meg szabályokat, amelyeket metaszabálynak neveztek el. Például az N_1 + de + N_2<$Lemma_2$> előfordulhat N_1 + A_1<$Lemma_2$> alakban is, ahol az azonos indexű elemek azonos karaktersorozatot jelölnek, így a terminusvariánsban a főnévi fejnek (N_1) meg kell egyeznie az eredeti terminus főnévi fejével, és a főnevet követő melléknév lemmájának meg kell egyeznie az eredeti terminus bővítményében szereplő főnév lemmájával. Ezen szabály alapján a *pression de sang* 'vérnyomás' terminusvariánsa a *pression sanguine* 'vérnyomás' (Jacquemin 2001).

A terminusjelöltek és azok variánsainak kinyeréséhez az annotált korpuszt használja fel a program. Ez alapján a bemenet alapján unifikációs nyelvtannal felismeri a terminusokat és a nyelvtan segítségével egyben szintaktikailag is elemzi. Ezek után végzi el a terminusjelöltek variánsainak keresését, amelyekhez a metaszabályokat, és egyéb tanult szabályokat alkalmazza (Jacquemin 2001).

A YaTeA program a TE-feladatot három lépésben oldja meg. A korpuszt először maximális főnévi csoportokra bontja, amit belső szintaktikai minták és határolóegységek alapján végez el (a ragozott ige például nem lehet terminus része, de terminus jobb vagy bal oldalán állhat határolóelemként). A határolóegységek listájához nem csak szófaji címkéket használ, hanem előre megadott elemeket is: a prepozíciók az angolban határolóegységek (pl. *without* 'nélkül'), de bizonyos prepozíciók nem: az *of* például lehet terminus része. A terminusjelölteket eztán szintaktikailag is elemzi: minden egyes terminusjelölthöz megadja annak belső szerkezetét is, azaz beazonosítja a fejet, és megadja rekurzívan annak módosítóit, és ez utóbbiak bővítményeit. A harmadik lépés statisztikai mértéket ad minden egyes terminusjelölthöz, amely a terminusjelöltek kézi validálásánál nyújt segítséget, ugyanis a nagyobb súllyal rendelkező jelöltek nagyobb valószínűséggel terminusok (Aubin és Hamon 2006).

[31] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

8.8. táblázat: A Fastr, a YaTeA és a saját terminológiakivonatoló eredményeinek összehasonlítása a szabadalmi korpuszon

Alkalmazás	A23L			G06F		
YaTeA	0,5711	0,3451	0,4270	0,5826	0,3045	0,3983
FastR	0,5764	0,4130	0,4806	0,5349	0,3962	0,4523
Saját	**0,7614**	**0,6176**	**0,6809**	**0,8280**	**0,6367**	**0,7185**

A 8.8. táblázat eredményeiből jól látható, hogy mindkét korpuszon a saját alkalmazás eredményei mind fedés, mind pontosság tekintetében meghaladták a másik kettő alkalmazás eredményeit. Bár mindhárom alkalmazás szabály alapon működik, az eredmények azért is lehetnek különbözőek, mert nem ugyanazt a morfoszintaktikai elemzőt használják.

A 8.9. táblázat az A23L korpusz különböző dokumentumain mért fedés, pontosság és F-értékeket mutatják.

8.9. táblázat: Az A23L korpuszon elért fedés, pontosság és F-értékek a három terminológiakivonatoló esetén

Dokumentum	TE-eszköz	fedés	pontosság	F-érték
EP2008056887	YaTeA	0,5380	0,3390	0,4160
	FastR	0,5556	0,3285	0,4128
	Saját	**0,7826**	**0,5806**	**0,6667**
EP2008061497	YaTeA	0,6076	0,3462	0,4410
	FastR	0,5149	0,3833	0,4395
	Saját	**0,7270**	**0,4977**	**0,5909**
EP2008063597	YaTeA	0,5176	0,3196	0,3952
	FastR	0,5246	0,3556	0,4238
	Saját	**0,7852**	**0,6935**	**0,7365**
EP2009057808	YaTeA	0,6086	0,2795	0,3830
	FastR	0,5631	0,4028	0,4696
	Saját	**0,7963**	**0,6099**	**0,6908**
FR2006001856	YaTeA	0,4769	0,4218	0,4477
	FastR	0,45	0,36	0,4
	Saját	**0,7848**	**0,6327**	**0,7006**
FR2007001526	YaTeA	0,5691	0,3249	0,4137
	FastR	0,6019	0,4218	0,496
	Saját	**0,7308**	**0,6835**	**0,7063**
FR2007051158	YaTeA	0,5119	0,3538	0,4184
	FastR	0,5919	0,4552	0,5146
	Saját	**0,7466**	**0,6301**	**0,6834**
FR2007051178	YaTeA	0,6438	0,4215	0,5095
	FastR	0,6716	0,5	0,5732
	Saját	**0,7705**	**0,6138**	**0,6833**
FR2007051372	YaTeA	0,6360	0,3005	0,4082
	FastR	0,6596	0,4794	0,5552
	Saját	**0,7608**	**0,6178**	**0,6819**
FR2007051549	YaTeA	0,6019	0,3439	0,4377
	FastR	0,6303	0,4438	0,5208
	Saját	**0,7293**	**0,6168**	**0,6684**
Átlag	YaTeA	0,5711	0,3451	0,4270
	FastR	0,5764	0,4130	0,4806
	Saját	**0,7614**	**0,6176**	**0,6809**

9. Összegzés és további kutatási lehetőségek

Az automatikus terminológiakivonatolás a számítógépes nyelvészeten és a terminográfián belül a számítógépes terminológia egy igen kutatott területe. Alkalmazása elég sokrétű: az automatikus szövegindexeléstől a terminológiai adatbázisok létrehozásán keresztül a fordítói munka elősegítéséig számos területen használatos. A feladat összetettségét jelzi már a terminusok definícióinak nagy száma, amelyek ráadásul az alkalmazási területtől is függnek. Célunk egy olyan terminológiakivonatoló létrehozása volt, amely egy adott, francia nyelvű szabadalmi leírásból kinyeri az összes terminust, amely valamilyen szaknyelvhez kötődik, és egy adott fogalmat denotál ezen a szaknyelven belül.

A disszertáció első három fejezete foglalkozott a disszertációhoz kapcsolódó fogalmak definiálásával. A negyedik fejezet célja a francia nyelvű főnévi egységek terminusokkal történő összevetése volt. Az ezt követő fejezetben fejtettük ki részletesen az automatikus terminológiakivonatolás lépéseit és a korábbi, valamint aktuális módszereit, főleg a nemzetközi szakirodalom alapján. A hatodik fejezetben írtuk le a használt korpuszt: 10-10 szabadalmi leírást választottunk az informatika, valamint az emberi szükségletek területéről. A hetedik fejezetben fejtettük ki a saját terminológiakivonatoló lépéseit, működését. A nyolcadik fejezetben ismertettük a kapott eredményeket, külön megvizsgálva, hogy az egyes módszerek milyen mértékben járultak hozzá a terminuskinyerés hatékonyságának javulásához.

Magyarországon a TE nem egy igen kutatott terület, a nemzetközi szakirodalomban annál inkább, főleg az angol nyelvre vonatkozóan. Disszertációnk egyik célja az volt, hogy hozzájáruljunk az itthoni terminuskinyerés kutatásához, különösen a francia nyelvre vonatkozóan. A francia nyelv választását az igazolta, hogy egyrészről a szakirodalomban a francia nyelvet tekintve kevesebb a terminológiakivonatolási kutatás, mint az angol nyelvre, másrészt azt, hogy ebben a nyelvben az összetett terminusok képzésekor gyakran alkalmaznak olyan összetettszó-képzési technikákat, amelyek elősegítik azok elkülönülését a köznyelvi egységektől, például a három egymást követő főnév (mint a *programmation côté serveur* 'szerver oldali programozás') nagy valószínűséggel terminus.

A TE általában hibrid módszerrel történik, tehát alkalmaznak szabály alapú és statisztikai módszereket is, amelyek közül az utóbbit inkább az első fázisban, azaz a terminuskinyerésben, az előbbit a kinyert terminusjelöltek szűrésében, azaz a második fázisban. A vizsgálatunk egyik célja az volt, hogy megmutassuk, a nemzetközi

szakirodalomban ezen általános nézőpont fordítottja is megfelelő eljárás lehet: a terminusjelölteket szabály alapon véges állapotú determinisztikus automatával nyertük ki, majd azokat statisztikai módszerekkel szűrtük.

Ezen eljárás esetén a szakirodalom alapján az előfeltételezésünk az volt, hogy a szabály alapú terminuskinyerés nagy fedést hoz alacsony pontossággal. Az eredmények ezt igazolták: a két típusú korpuszban a terminuskinyerési szakaszban a fedés 0,8, a pontosság 0,5 körüli értékű volt. A saját terminológiakivonatoló a szabály alapú kinyerés előtt a szöveget szabály alapon szűrte is. Ehhez létrehoztunk egy *stopword*-listát, amely a lehető legtöbb, főnévvel rendelkező konnektívumot tartalmazta, mint a *par exemple* 'például' vagy az *en effet* 'ugyanis'. Ezen kívül olyan melléknevek, határozószók is szerepeltek ezen listában, amelyek nem lehetnek terminusok részei, mint a *suivant* 'következő' melléknév vagy a *très* 'nagyon' határozószó. A fentieken kívül a listába még belevettük a vonzattal együtt megjelenő mellékneveket is, mert azok sem lehetnek terminusok részei, például az *apte à* 'valamire alkalmas'.

A kutatásunk második célja az volt, hogy megtudjuk, hogy a szabály alapú szűrés milyen mértékben növeli a pontosságot. Ezen szűrés után a fedés értékek kis mértékben, a pontosság adatai jelentősen, körülbelül 0,09-dal emelkedtek, 0,6 körüli értékre.

A szabály alapú szűrést és kinyerést követte a statisztikai alapú szűrés, amelyhez három statisztikai mértéket használtunk. Az egyik egy *termhood*-érték, a *weirdness*, amely azt mutatja meg, hogy az adott terminusjelölt milyen arányban fordul elő a szaknyelvben egy köznyelvi korpuszhoz viszonyítva. A második a C-érték, amely azt mutatja meg, hogy az adott többszavas terminusjelölt elemei mennyire tartoznak egybe, azaz azt mondja meg, hogy az a terminusjelölt teljes mértékben terminusjelölt-e, vagy inkább csak egy része terminus, vagy egy nagyobb terminus része. A harmadik a súlyérték, amely azt határozza meg, hogy a terminusjelölt előtti és utáni szavak, írásjelek milyen valószínűséggel fordulnak elő terminusok mellett, ezáltal azt is meg lehet becsülni, hogy az adott terminusjelölt a szövegkörnyezete alapján milyen valószínűséggel terminus. Ezen három mértékből minden terminusjelöltre egy összevont értéket alkalmaztunk, amely a [0..1] irracionális tartományban helyezkedett el.

Előfeltételezésünk az volt, hogy az egyszavas terminusoknál a *weirdness*-érték határozza meg nagyobb mértékben azt, hogy az tényleg terminus-e, a többszavasaknál pedig inkább a súly- és a C-érték, mivel az összetett szavakra internetes keresőmotoron történő keresés csak megközelítő értéket adhat. Ezen hipotézisünket az eredmények

részben igazolták: az első öt leírásra lefuttatott program azt mutatta, hogy az egyszavas terminusoknál a *weirdness*-érték tényleg meghatározó (kb. 80%-ban), a súlyérték kevésbé (kb. 20%-ban). A többszavas terminusok esetében a C-érték nem hozott javulást, mert a legjobb eredmények akkor születtek, ha azt 0%-ban vettük figyelembe. Így a többszavas terminusoknál is a súly- és *weirdness*-értékeket alkalmaztuk.

A statisztikai mértékek használatánál az volt az egyik előfeltételezésünk, hogy amennyiben az elfogadási küszöböt nagynak állítjuk be, akkor a pontosság jelentős mértékben nő a fedés jelentős csökkenésével. Ez így is történt: a statisztikai módszerekkel átlagosan 0,9-es pontosságot is el lehet érni, de akkor a fedés 0,02 érték körüli.

A másik hipotézisünk az volt, hogy ha megfelelő küszöbértéket állítunk be, akkor az összevont statisztikai érték a pontosságot jelentősen megnöveli a fedés minél kisebb csökkenésével, azaz jelentősen meg tudja növelni a két mértékből kiszámítható F-értéket. Ez az előfeltételezésünk nem igazolódott, mert a legnagyobb F-értéket akkor tudtuk elérni, ha a küszöbértéket alacsonynak választottuk, ekkor viszont ez a növekedés a kétféle korpuszon átlagban mindössze 1,5%-os.

A terminológiakivonatoló alkalmazás eredményeit más TE-feladatot megvalósító alkalmazással is összehasonlítottuk, mégpedig a Fastr és a YaTeA programmal. Az utóbbi két program esetén a fedés 57%-os, a pontosság 35% körüli értékeket mutatott, amely alulmaradt az általunk létrehozott terminológiakivonatoló eredményein. Ez az eredmény azonban annak is betudható, hogy a két külső programhoz a *TreeTagger* nevű *POS-tagger*t használtuk, mert ez utóbbiak azt az annotációs formátumot támogatják.

Az eredmények legfontosabb üzenete, hogy a terminusjelölt-lista létrehozása nemcsak a széles körben alkalmazott statisztikai módszerekkel, hanem szabály alapú, azaz nyelvészeti módszerekkel is lehet eredményes francia nyelvű szabadalmakon, különösen a fedés tekintetében.

A kutatást a későbbiekben érdemes lehet kiterjeszteni más típusú korpuszokra is, nemcsak szabadalmakra, mert ez utóbbiakban jelentősebb arányban fordulnak elő terminusok, hanem más, didaktikai jellegűbb, általánosabb szövegekre. Az általánosabb szakmai szövegeknek (például a didaktikusabb, magyarázó jellegű leírásoknak) megvannak az előnyei és a hátrányai is. Egy általánosabb szöveg esetében hátrány, hogy nemcsak terminusokat használ a szövegben, hanem sok olyan elem is előfordul benne, ami nem terminus, azért hogy a szöveg ne legyen olyan száraz, mint mondjuk egy szabadalmi leírás. Ennek következtében a statisztikai szűrőknek nagyobb szerepet kell kapniuk, hogy a

nem terminus elemeket is kiszűrjük. Előnyük viszont az, hogy az új fogalmak valamilyen módon definiálva is vannak. A disszertációban csak említettem, hogy léteznek olyan algoritmusok, amelyek terminusok és a hozzájuk tartozó definíciók kinyerésére alkalmasak, de nem fejtettem ki bővebben, mert a szabadalmakra a definiálások kevésbé jellemzők. Didaktikusabb szövegekben feltételezhetjük, hogy vagy alapterminusról van szó, vagy a szövegben definiálva van, és akkor az 5.3.4. fejezetben leírt, definíciók alapján történő kinyerést is el lehet végezni.

Ezenkívül érdemes lenne megnézni, hogy a terminusok statisztikai szűrése történhetne-e más hasonló mértékekkel is, amelyek részletes listája az 5. fejezetben olvasható. Ebben az esetben lehet, hogy a C-érték helyett más *unithood*-mértékkel jobb eredményt érhetünk el, és ezáltal ki tudjuk szűrni azokat a többszavas terminusokat, amelyekben például a terminus részét nem képző melléknévi utómódosító található.

A 10-10 szabadalmi leírásból érdemes lenne egy olyan korpuszt összeállítani, ahol a terminusok nem külön listában szerepelnek, hanem a szövegkörnyezettel együtt, az eredeti szövegben bejelölve. Erre azért lenne szükség, mert egy ilyen korpusz már tanuló algoritmusoknak is alapul szolgálhat, és jelenleg nem érhető el terminológiai korpusz, csak terminológiai adatbázis, ahol a terminusok csak fel vannak sorolva.

A disszertáció keretében létrehozott terminológiakivonatoló csak francia nyelvű szabadalmi szövegeken működik, más nyelvekre való kiterjesztését több tényező is gátolja. Egyrészt, a TE-alkalmazás szabály alapú szűrést végez a szövegen, szabály alapú módszerrel nyeri ki a terminusokat, majd ezekre alkalmaz statisztikai szűrőket. A szabály alapú kinyeréshez a szöveget mindenképpen szótövesíteni kell, és a benne található szavakhoz hozzá kell rendelni egy szófaji kategóriát. Ha olyan nyelvet választunk, amihez létezik ilyen, márpedig a legtöbb nyelvhez van, akkor szabály alapú módszereket is alkalmazhatunk. A szabály alapú kinyeréshez azonban mintákat kell megadni, amik mindenképpen nyelvfüggőek. Ezeket nehéz lenne nyelvfüggetleníteni, mert például a francia terminusokban gyakori prepozíció nincs is a magyarban.

A szabály alapú kinyerést ekkor nem pontos mintákkal hajthatnánk végre, hanem határolókkal (Bourigault 1994), azaz fordítottan járhatunk el: nem a terminusokra jellemző mintákat nyerünk ki, hanem azon tokeneket szűrjük ki, amik biztos nem terminusok részei, például vessző, igék stb. A szabály alapú szűrésnél előre megadott elemeket szűrünk ki a listából. Ezek legtöbbjének konkrét megvalósulása (szóalakja) nyelvfüggő, de a kifejezések megfeleltethetőek egymásnak. Ha egy szótárból ezen szavakat, kifejezéseket ki

tudjuk nyerni, ez is megvalósítható lenne nyelvfüggetlen alapon. Ehhez viszont hozzátartozik, hogy vannak olyan kifejezések, amelyeket nem kellene minden nyelvben kiszűrni. Például a franciában az *ugyanis* megfelelője az *en effet*, egy prepozícióból és főnévből áll, ami tipikusan terminus-összetétel, de mégsem lehet terminus része, így azt ki kell szűrni. A magyar változata, amely kötőszóként használatos, nem is lehetne terminus része, így kiszűrni sem kell a szövegből.

Egyedül a statisztikai mértékek azok, amelyek nyelvfüggetlenek. Az első, a *termhood*-érték, az adott terminusjelölt köznyelvi és szaknyelvi előfordulásának aránya. Ha egy jelölt a szaknyelvben fordul elő gyakrabban, az terminus: ez minden nyelvre igaz, de ezen értékhez szükséges, hogy az adott nyelvre legyen egy köznyelvi referenciakorpusz. A *unithood*-mérték adja meg a terminusok részelemeinek kohéziós értékét, és ez is bármely nyelvre alkalmazható. Egyedül a súlyérték az, (amely a szavak szövegkörnyezete alapján mond egy statisztikai értéket) amelyhez az adott nyelvre vonatkozóan tanulókorpusszal kell rendelkeznünk, mert ez is valószínűleg minden nyelvben más, például vannak olyan nyelvek, ahol nincsenek névelők, és akkor más veszi át a szerepüket.

Kis B. (2005) alapján, például ha egy könyvből kell TE-t végezni, akkor a minták gépi tanulásához egy adott nyelvre fel lehet használni a könyv tárgymutatóját is, amely főleg terminusokat tartalmaz. Abból a program automatikusan ki tudja nyerni a jellemző mintákat és a szövegkörnyezetet. Valamint statisztikát tud készíteni arról, hogy például mi fordul elő szövegkörnyezetükben, vagy hogy abban a nyelvben mik a jellemző terminusminták.

A lehetséges nyelvfüggetlenítés során mind a pontosság, mind a fedés jelentősen csökkenne, hiszen nem lehet nyelvspecifikus szabályokat megadni, amelyek hozzájárulnának az alkalmazás jelentős mértékű hatékonyságához.

Köszönetnyilvánítás

Köszönöm témavezetőmnek, dr. Váradi Tamásnak a disszertáció írásában nyújtott hasznos tanácsait.

Köszönöm a Nyelvtudományi Doktori Iskola Francia nyelvészet szakirány vezetőjének, dr. Gécseg Zsuzsannának, aki másodtémavezetőként segített a doktori disszertáció egyes fejezeteinek átnézésében, és aki mindig emlékeztetett a határidőkre. Köszönöm még a Francia Nyelvi és Irodalmi Tanszék vezetőjének, dr. Szász Gézának, aki sok esetben megkönnyítette az egyetemi bürokráciában való kiigazodást, valamint hogy lehetőséget adott arra, hogy a Tanszék alkalmazottjaként a doktori disszertációm megírásához a körülmények adottak legyenek.

Köszönöm a Szegedi Tudományegyetem Nyelvtudományi Doktori Iskola oktatóinak a támogatását és a magas színvonalú képzést. Hálás vagyok dr. Kenesei Istvánnak a doktori képzés alatt nyújtott segítségéért.

Köszönettel tartozom dr. Csirik Jánosnak és dr. Alexin Zoltánnak a lehetőségért, hogy a Szegedi Tudományegyetem Informatikai Tanszékcsoportján a MaSzeKer-projekt keretében elmélyültebben foglalkozhattam a szabadalmak nyelvészeti elemzésével, valamint átláthattam egy komplex szintaktikai elemző működését.

Végül köszönöm dr. Lászlónak az alapos nyelvi korrektúrát és hasznos tanácsait.

Források

Csábi, Sz. (szerk.), 2007, *Magyar értelmező kéziszótár*, Budapest, Akadémiai Kiadó.

Iványi, A. (szerk.), 2006, *Angol–magyar informatikai szótár*, Budapest, Tinta Könyvkiadó.

Laczkó, K., Mártonfi, A., 2004, *Helyesírás*, Budapest, Osiris Kiadó.

(*Le Nouveau*) *Petit Robert 2010*. On-line kiadás. http://portail.cns-edu.com/

Les marqueurs de relation, http://www.colvir.net/prof/michel.durand/marqueurs.html.

Várlaki, T. (szerk.), 2005, *Számítástechnikai és informatikai szakkifejezések 11 nyelven*, Budapest, Akkord Könyvkiadó.

Vizi, E. Sz. (szerk.), 2003, *Magyar nagylexikon* (16. és 17. kötet). Budapest, Magyar Nagylexikon Kiadó.

WIPO Patentscope, http://www.wipo.int/pctdb/en/ (szabadalmi kereső, az idézett szabadalmak forrása)

Bibliográfia

Abeillé, A., Godard, D. (1999). La position de l'adjectif en français: le poids des mots. *Recherches Linguistiques de Vincennes* **28**: 9–31.

Adjiman, Ph. (2009). Open Calais From Java: Get Ready To Extract Entities, Facts And Events In 4 Minutes!, http://philippeadjiman.com/blog/2009/09/16/open-calais-from-java-with-eclipse-extract-entities-facts-and-events-in-4-minutes/, informatikai blog, poszt időpontja: 2009. szeptember 16.

Ahmad, K., Gillam, L., Tostevin, L. (1999). University of surrey participation in trec8: Weirdness indexing for logical document extrapolation and retrieval (wilder). In *The Eighth Text REtrieval Conference (TREC-8)*.

Ahmad, K. (2001). The role of specialist terminology in artificial intelligence and knowledge acquisition. In Wright, S-E., Budin, G. (eds.) *Handbook of terminology management 2*, 809–844.

Allen, G., Owens, M. (2010). *The Definitive Guide to SQLite* (2. kiadás). New York (NY), Apress.

Angster, E. (2004). *Objektumorientált programozás, Java 2*. Budapest, 4körBt.

Anscombre, M. J-C. (1990). Article zéro et structuration d'événements. In Charolles, M., Jayez, J. (eds.) *Le discours: représentations et interprétations*. P.U.N., 265–300.

Anscombre, M. J-C. (1991). L'article zéro sous préposition, *Langue française* **91**: 24–39.

Anstein, S., Kremer, G., Reyle, U. (2006). Identifying and Classifying Terms in the Life Sciences: The Case of Chemical Terminology, *Proceedings of LREC 2006* (CD-ROM), Genova, 1095–1098.

Aubin, S., Hamon, Th. (2006). Improving term extraction with terminological resources. *Advances in Natural Language Processing Lecture Notes in Computer Science*. 380–387.

Auger, P. (1988). La terminologie au Québec et dans le monde, de la naissance à la maturité. In Gaumond, J-C. (ed.) *L'ère nouvelle de la terminologie. Actes du 6e colloque OLF-STQ de terminologie*. Montréal, OLF, 27–59.

Banay, G.L. (1948). *An Introduction to Medical Terminology, I. Greek and Latin Derivations*. Bulletin medical library Association, 1–27.

Basili, R., Moschitti, A., Pazienza, M., Zanzotto, F. (2001). A contrastive approach to term extraction. In *Proceedings of the 4th Terminology and Artifcial Intelligence Conference (TIA)*. France.

Beaugrande, Robert de. (1987). Special purpose language and linguistic theory. *ALSED-LSP Newsletter* **10**, 2(25): 2–10.

Béjoint, H., Ahronian, C. (2008). Les noms composés anglais et français du domaine d'Internet: une radiographie bilingue, *Meta: journal des traducteurs* **53**(3): 648–666.

Benajiba, Y. (2009). *Named entity recognition*. PhD-értekezés, Universidad Politécnica de Valencia, May, http://users.dsic.upv.es/~prosso/resources/BenajibaPhD.pdf.

Bergenholtz, H., Kaufmann, U. (1997). Terminography and lexicography. A critical survey of dictionaries from a single specialised field. *Hermes* **18**: 91–125.

Bessé, B. de, Nkwenti-Azeh, B., Sager, J. C. (1997). Glossary of terms used in Terminology. *Terminology* **4**(1): 117–156.

Bikel, D. M., Miller, S., Schwartz, R., Weischedel, R. (1997). Nymble: a high-performance learning name finder, *Proceedings of the 5th Conference on Applied Natural Language Processing*. 194–201.

Bosredon B., Tamba I. (1991). *Verre à pied, moule à gaufres*: préposition et noms composés de sous-classe. *Langue française* **91**: 40–55.
Bouchard, D. (1998). The distribution and interpretation of adjectives in French: a consequence of Base Phrase Structure. *Probus* **10**: 139–183.
Boulaknadel, S., Daille, B., Aboutajdine, D. (2008). A Multi-Word Term Extraction Program for Arabic Language, *Proceedings of LREC 2008* (CD-ROM), Marrakech, 1485–1488.
Bourigault, D. (1994). *Lexter, un logiciel d'extraction de terminologie. Application à l'acquisition des connaissances à partir de textes*, Doktori disszertáció, Informatique Appliquées aux Sciences Humaines de l'Ecole des Hautes Etudes en Sciences Sociales, Paris.
Brekke, M., Innselset, K., Kristiansen, M., Øvsthus, K. (2006). Automatic Term Extraction from Knowledge Bank of Economics, *Proceedings of LREC 2006* (CD-ROM). Genova, 1912–1916.
Cabré, M. T. (1999). *Terminology. Theory, methods and applications*, Amsterdam/Philadelphia, John Benjamins.
Cabré, M. T., Bagot, R. E., Vivaldi Palatresi, J. (2001). Automatic term detection. A review of current systems. In Bourrigault, D., Jacquemin, Ch., L'Homme, M-C. (eds.) *Recent advantages in Computational Terminology*. Amsterdam/Philadelphia, John Benjamins Publishing Co., 53–87.
Cabré, M. T. (2003). Theories of terminology. Their description, prescription and explanation. *Terminology* **9**(2): 163–200.
Cadiot, P. (1993). *À* entre deux noms: vers la composition nominale. *Lexique* **11**: 193–240.
Carroll, J. (2003). Parsing, In Mitkov, R. (ed.) *The Oxford Handbook of Computational Linguistics*. Oxford, Oxford University Press, 233–248.
Church, K., Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics* **16**(1), 22–29.
Cinque, G. (1994). On the evidence for partial N-movement in the Romance DP. In Cinque, G. (ed.) *Path Towards Universal Grammar. Studies in Honor of Richard Kayne*. Washington, Georgetown University Press, 85–110.
ClearForest Gnosis (2009). https://addons.mozilla.org/hu/firefox/addon/3999/, Firefox-kiegészítőket tartalmazó oldal.
Cohen, J. D. (1995). Highlights: Language- and domainindependent automatic indexing terms for abstracting. *Journal of the American Society for Information Science* **46**(3): 162–174.
Cormen, T. H., Leiserson, C. E., Rivest R. L., Stein, C. (2003). *Új algoritmusok*. Budapest, Scolar Kft.
Daille, B. (1994). *Approche mixte pour l'extraction automatique de terminologie: statistiques lexicales et filtres linguistiques*. PhD-értkezés. Université de Paris VII, Paris.
Daille, B. (1995). Study and implementation of combined techniques for automatic extraction of terminology. In Klavans, J. and Resnik, P. (eds.) *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*. Cambridge, Massachusetts; London, England MITPress, 49–66.
Daille, B. (2005). Variations and application-oriented terminology engineering. *International journal of theoretical and applied issues in specialized communication* **11**(1): 181–197.
Deane, P. (2005). A nonparametric method for extraction of candidate phrasal terms. In *Proceedings of the Conference 43rd Annual Meeting of the Association for*

Computational Linguistics. University of Michigan, USA: The Association for Computer Linguistics, 605–613.
Depecker, L. (2000). Le signe entre signifié et concept. In Béjoint, H., Thoiron, Ph. (eds.) *Le sens en terminologie*. Lyon, Presses Universitaires de Lyon, 86–126.
Dias, G., Kaalep, H. (2003). Automatic Extraction of Multiword Units for Estonian: Phrasal Verbs. *Languages in Development* **41**: 81–91.
Drouin, P. (2003). Term extraction using non-technical corpora as a point of leverage. *Terminology* **9**(1): 99–115.
Enguehard, Ch. (2005). Un banc de test pour la reconnaissance de termes en corpus. In Williams, G.. (ed.) *La linguistique de corpus*. Rennes, Presses Universiataires de Rennes, 273–286.
Evans, D. A., Lefferts, R. G. (1995). Clarit-trec experiments. *Information Processing and Management* **1**(3): 385–395.
Fahmi, I., Bouma, G., Plas, L. van der. (2007). Using Multilingual Terms for Biomedical Term Extraction. In *Proceedings of the RANLP Workshop on Acquisition and Management of Multilingual Lexicons*. Borovetz, Bulgaria, 1–8.
Foo, J. (2009). Term extraction using machine learning. *Proceedings of NODALIDA 2007 (CD-ROM)*, 349–357.
Fóris, Á. (2005). *Hat terminológiai lecke*. Pécs, Lexikográfia Kiadó.
Fóris, Á. (2007). A terminusok és a terminológia rendszer. In Heltai, P. (szerk) *Nyelvi Modernizáció. Szaknyelv, fordítás, terminológia. XVI. Magyar Alkalmazott Nyelvészeti Kongresszus előadásai* (Gödöllő, 2006. április 10–12). Pécs – Gödöllő, MANYE – Szent István Egyetem, 15–26.
Fóris, Á., Kékesi, N., Kozma, L. (2009). Megjegyzések a terminusok jelentésmeghatározásának módszeréhez. *Magyar Terminológia* **2**: 99–112.
Frantzi, K. T., Ananiadou, S. (1999). The c/nc value domain independent method for multi-word term extraction. *Journal of Natural Language Processing* **6**(3): 145–179.
Frantzi, K.T., Ananiadou, S. (1997). Automatic term recognition using contextual clues. In *Proceedings of Mulsaic 97, IJCAI*. Japan, Japan, 1997.
Frantzi, K.T., Ananiadou, S., Tsujii, J. (1998). The C-Value/NC-Value Method of Automatic Recognition for Multi-word terms. In *ECDL '98 Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries*, 585–604.
Fülöp, Z. (2004). *Formális nyelvek és szintaktikai elemzésük*, Szeged, Polygon.
Galinski, C., Nedobity, W. (1988). Special languages, terminology planning and standardization. In Strechlow, R. E. (ed.) *Standardization of technical terminology. Principles and practices*. Baltimore, American Society for Testing and Materials. 4–12.
Garcia-Molina, H., Ullman, J.D., Widom, J. (2002). *Database Systems: The Complete Book*. Prentice Hall.
Gelboukh, A., Sidorov, G., Lavin-Villa, E., Chanona-Hernandez, L. (2010). Automatic Term Extraction using Log-likelihood based Comparison with General Reference Corpus. In Hopfe, J. és mtsai (eds.) *Natural Language processing and information systems*. *Proceedings of NLDB 2010*. Berlin, Springer-Verlag, 248–255.
Gotti, M. (2003). *Specialized Discourse. Linguistic Features and Changing Conventions*, Bern, Peter Lang.
Grishman, R. (2003). Information Extraction. In Mitkov, R. (ed.) *The Oxford Handbook of Computational Linguistics*. Oxford, Oxford University Press, 545–559.

Gross, M. (1996). *Les expressions figées en français: noms composés et autres locutions*. Paris, Ophris.

Ha, L.A., Fernandez, G., Mitkov, R., Corpas, G. (2008). Mutual bilingual terminology extraction. In *Proceedings of LREC 2008* (CD-ROM). Marrakech, 1818–1824.

Hoste, V., Lefever, E., Vanopstal, K., Delaere, I. (2008). Learning-based Detection of Scientific Terms in Patient Information. In *Proceedings of LREC 2008* (CD-ROM), Marrakech, 585–591.

Jacquemin, Ch., Bourrigault, D. (2003). Term extraction and automatic indexing. In Mitkov, R. (ed.) *The Oxford handbook of computational linguistics*. Oxford, Oxford University Press, 599–615.

Jacquemin, Ch. (2001). *Spotting and discovering terms through natural language processing*. Cambridge, Cambridge(MA)/London, MIT Press.

Jakobson, R. (1963). Essais de linguistique générale. Paris, Éditions de minuit.

JFLAP [Java Formal Languages and Automata Package] http://jflap.org

Jobst, Á. (2007). A mi mint a hatalom és a szolidaritás névmása. *Magyar nyelvőr* **131**(1): 29–47.

Kelemen, J. (2001). *Grammaire du français contemporain*. Budapest, Nemzeti Tankönyvkiadó.

Kis, Á. (2007). Automatikus terminuskivonatolás diszkurzusszerkezetek segítségével. In Pusztay, J. (szerk.) *A magyar, mint veszélyeztetett nyelv?* Szombathely, BDF. 165–181.

Kis, B. (2005). Automatikus terminológia keresés számítógéppel – kísérlet, *Fordítástudomány* **7**(1): 84–96.

Kiss, J. (1995). Társadalom és nyelvhasználat. Budapest, Nemzetközi Tankönyvkiadó.

Kocourek, R. (1982). *La langue française de la technique et de la science*. Wiesbaden, Brandstetter.

Kovács, F. (2001). *A magyar nyelvtudományi terminológia kialakulása*. Budapest, Akadémai Kiadó.

Kreibich, J. A. (2010). *Using SQLite*. Sebastopol (CA), O'Reilly Media.

Kurz, D., Xu, F. (2002). Text mining for the extraction of domain relevant terms and term collocations. In *Proceedings of the International Workshop on Computational Approaches to Collocations*. Vienna.

Laenzlinger, C. (2003). *Initiation à la Syntaxe formelle du français: Le modèle Principes et Paramètres de la Grammaire Générative Transformationnelle*. Peter Lang AG, Berne.

Lefever, E., Macken, L., Hoste, V. (2009). Language-independent bilingual terminology extraction from a multilingual parallel corpus. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*. Athens, 409–504.

Lérat, P. (1989). Les *fondements théoriques* de la terminologie. *La banque des mots* (különszám): 51–62.

L'Homme, M-C. (2004). *La terminologie: principes et techniques*. Montréal, Les Presses de l'Université de Montréal.

Macken, L., Lefever, E., Hoste, V. (2008). Linguistically-based sub-sentential alignment for terminology extraction from a bilingual automotive corpus. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*. Manchester, 529–536.

Manning, Ch. D., Raghavan, P., Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge, Cambridge University Press.

Martín-Vide, C. (2003). Formal Grammars and Languages. In Mitkov, R. (ed.) *The Oxford Handbook of Computational Linguistics*. Oxford, Oxford University Press, 157–177.

Mathieu-Colas, M. (1996). Essai de typologie des noms composés français. *Cahiers de lexicologie* **69**: 71–125.

Maynard, D., Ananiadou, S. (2000). Identifying Terms by their Family and Friends. In *Proceedings of COLING 2000*. Luxembourg, 530–536.

Meilland, J.-C., Bellot, P. (2005). Extraction automatique de terminologie à partir de libellés textuels courts. In Williams, G. (ed.) *La linguistique de corpus*. Rennes, Presses Universitaires de Rennes, 357–370.

Mikheev, A., Grover, C., Moens, M. (1998). Description of the LTG system used for MUC-7, Proceedings of the 7th Message Understanding Conference (MUC), http://www-nlpir.nist.gov/related_projects/muc/proceedings/muc_7_proceedings/ltg_muc7.pdf

Mikheev, A. (2003). Text Segmentation. In Mitkov, R. (ed.) *The Oxford Handbook of Computational Linguistics*. Oxford, Oxford University Press, 201–218.

Milios, E., Zhang, Y., He, B., Dong, L. (2003). Automatic Term Extraction and Document Similarity in Special Text Corpora. In *Proceedings of the 6th Conference of the Pacific Association for Computational Linguistics (PACLing'03)*. Halifax, Nova Scotia, Canada, 275–284.

Mima, H., Ananiadou, S. (2001). An Application and Evaluation of the C/NC-value Approach for the Automatic term Recognition of Multi-Word units in Japanese. *International Journal on Terminology* **6**(2): 175–194.

Murphy, M. L. (2010). *Beginning Android 2*. New York, Apress.

Nadeau, D. (2007). *Semi-Supervised Named Entity Recognition: Learning to Recognize 100 Entity Types with Little Supervision*. Ottawa-Carleton Institute for Computer Science, University of Ottawa.

Nagy, Á. (2008). L'extraction terminologique: l'un des défis actuels de la linguistique informatique, In Kovács, K., Nagy, Á. (szerk.) *Le Passé dans le Présent, le Présent dans le Passé.* Séminaire doctoral, Szeged, 2007. október 25–26., Szeged, JATEPress. 283–290.

Nagy, Á. (2009a). Kísérlet szintaktikai és statisztikai módszerekkel történő automatikus terminológiakivonatolásra francia nyelvű szövegekből. In Sinkovics, B. (szerk.) *LingDok 8. Nyelvész-doktoranduszok dolgozatai*. JATEPress, 71–86.

Nagy, Á. (2009b). *Főnévi csoportok kinyerése szabály alapú és statisztikai módszerekkel*. szakdolgozat, Szegedi Tudományegyetem Informatikai Tanszékcsoport.

Nagy, Á. (2009c). La structure interne des termes techniques du français et leur reconnaissance par ordinateur. In Kieliszczyk, A., Pilecka, E. (eds.), *La perspective interdisciplinaire des études françaises et francophones*. Łask, Oficyna Wydawnicza LEKSEM, 117–123.

Nybakken, O.E. (1979). *Greek and Latin in scientific terminology*. Ames, The Iowa State University Press.

Osenga, K. (2006). Linguistics and patent claim construction. *Rutgers Law Journal* **38**(61): 61–108.

Otman, G. (1995). *Les représentations sémantiques en terminologie*. Doktori disszertáció, UFR Langue française Université Paris IV – Sorbonne.

Petit, G. (2001). L'introuvable identité du terme technique. *Revue Française de Linguistique Appliquée* **VI**(2): 63–79.

Picht, H., Draskau, J. (1985). *Terminology: An introduction*. Guilford, University of Surrey.

Piao, S., McNaught, J., Ananiadou, S. (2008). Clustering Related Terms with Definitions. *Proceedings of LREC 2008* (CD-ROM). Marrakech, 2013–2019.

Plante, P, Dumas, L. (1989). Le dépouillement terminologique assisté par ordinateur, *Terminogramme* **46**: 24–28.

Pukelsheim, F. (1994). The Three Sigma Rule. *The American Statistician* **48**(2): 88–91.

Pusztay, J. (2008). A terminológia mint a nyelv megmaradásának feltétele. Az oroszországi finnugor nyelvek helyzete. *Magyar Terminológia* **II/1**: 205–216.

Rayson, P., Garside, R. (2000). Comparing corpora using frequency profiling. In *Proceedings of the workshop on comparing corpora, 38th annual meeting of the Association for Computational Linguistics (ACL2000)*. 1–6.

Rey, A. (1979). *La terminologie: noms et notions*. Paris, Presses Universitaires de France.

Rey, A. (1995). *Essays in Terminology*. Amsterdam/Philadelphia, John Benjamins.

Riegel, M., Pellat, J-Ch., Rioul, R. (2009). *Grammaire méthodique du français* (4. kiadás). Paris, PUF.

Sager, J. C. (1990). *A practical course in terminology processing*. Amsterdam & Philadelphia, John Benjamins.

Sager, J. C. (2000). Pour une approche fonctionnelle de la terminologie. In Thoiron, Ph., Béjoint, H. (eds.) *Le sens en terminologie*. Lyon, Presses universitaires de Lyon, 40–60.

Sager, J. C., Dungworth, D., McDonald, P. F. (1980). *English Special languages. Principles and practice in science and technology*. Wiesbaden, Brandstetter.

Salton, G., Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management* **24**(5): 513–523.

Sauron, V. A. (2002). Tearing out the terms: evaluating terms extractors. In *Proceedings of Translating and the Computer 24*. London, Britain.

Saussure, F. de (1998). *Bevezetés az általános nyelvészetbe*. Budapest, Corvina Kiadó.

Savary, A. (2000). *Recensement et description des mots composés – méthodes et applications*. doktori disszertáció, Université de Marne-la-Vallée és Laboratoire d'Automatique Documentaire et Linguistique, Université Paris 7.

Sekine, S., Grishman, R., Shinnou, H. (1998). A decision tree method for finding and classifying names in Japanese texts. In *Proceedings of the 6th Workshop on Very Large Corpora*. 171–177.

Silberztein, M. (1990). Le dictionnaire électronique des mots composés. *Langue française* **87**: 71–83.

Slodzian, M. (2000). L'émergence d'une terminologie textuelle et le retour du sens. In : Béjoint, H., Thoiron, Ph. (eds.) *Le sens en terminologie*. Lyon, Presses Universitaires de Lyon, 61–85.

Schneuwly, B., Rosat, M-C., Dolz, J. (1989). Les organisateurs textuels dans quatre types de textes écrits. Etude chez des élèves de 10, 12 et 14 ans. *Langue française* **81**: 40–58.

Strevens, P. (1977). Special Purpose Language learning: a perspective. *Language Teaching & Linguistics Abstracts* **10**(3): 145–163.

Szalai, K., Ferenczhalmy, R., Fülöp, É., Vincze, O., László, J. (2009). Történelmi szövegek narratív pszichológiai vizsgálata a nemzeti identitás tükrében. In Tanács, A., Vincze, V. (szerk.) *MSZNY 2009*. Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged.

Váradi, T., Pintér, T., Mittelholcz, I., Peredy, M. (2010). Bibliográfiai hivatkozások automatikus kinyerése. In Tanács, A., Vincze, V. (szerk.) *MSZNY 2010*. Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged.

Vasconcellos, M. (2001). Terminology and Machine translation. In Wright, S-E., Budin, G. (eds.) *Handbook of terminology management 2*. Amsterdam/Philadelphia, John Benjamins, 697–723.

Vintar, Š. (2004). Comparative Evaluation of C-value in the Treatment of Nested Terms. Memura 2004 – Methodologies and Evaluation of Multiword Units in Real-World Applications. In *Proceedings of the International Conference on Language Resources and Evaluation 2004*. 54–57.

Vossen, P. (2003). Ontologies. In Mitkov, R. (ed.) *The Oxford Handbook of Computational Linguistics*. Oxford, Oxford University Press, 464–482.

Voutilainen, A. (2003). Part-of-Speech Tagging. In Mitkov, R. (ed.) *The Oxford Handbook of Computational Linguistics*. Oxford, Oxford University Press, 219–233.

Vu, T., Aw. A. T., Zhang, M. (2008). Term extraction through unithood and termhood unification. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP-08)*, Hyderabad (India), 631–636.

Wermter, J., Hahn, U. (2005). Finding new terminology in very large corpora. In Clark, P., Schreiber, G. (eds.) *Proceedings of the 3rd International Conference on Knowledge Capture (K-CAP 2005)*, Banff, Alberta, Canada: ACM, 137–144.

WIPO Intellectual Property Handbook: Policy, Law and use. (2004, 2. kiadás). Genf, WIPO.

Wong, W., Liu, W., Bennamoun, M. (2008). Determination of unithood and termhood for term extraction. In Song, M., Wu, Y. (eds.) *Handbook of Research on Text and Web Mining Technologies*, IGI Global.

Wüster E. (1931/1970). *Internationale Sprachnormung in der Technik: besonders in der Elektrotechnik*. Berlin, VDI-Verlag.

Wüster, E. (1976). La théorie générale de la terminologie, un domaine interdisciplinaire impliquant la linguistique, la logique, l'ontologie, l'informatique et la science des objets. *Actes du colloque international de terminologie*. Québec 5–8 octobre 1975, Québec, L'Éditeur officielle du Québec.

Wüster, E. (1981). L'étude scientifique générale de la terminologie, zone frontalière entre la linguistique, la logique, l'ontologie, l'informatique et les sciences des choses. In Rondeau, G., Felber, H. (eds.) *Textes choisis de terminologie. Vol. I: Fondements théoriques de la terminologie*. Québec, Université Laval – GIRSTERM, 55–114.

Yang, Y., Lu, Q., Zhao, T. (2008). Chinese Term Extraction Based on Delimiters. In *Proceedings of LREC 2008* (CD-ROM). Marrakech, 247–254.

Yirong, Ch., Qin, L., Wenjie, L., Zhifang, Sh., Luning, J. (2006). A Study on Terminology Extraction Based on Classified Corpora. In *Proceedings of LREC 2006* (CD-ROM). Genova, 2383–2386.

Yirong, Ch., Qin, L., Wenjie, L., Gaoying, C. (2008). Chinese Core Ontology Construction from a Bilingual Term Bank. *In Proceedings of LREC 2008* (CD-ROM). Marrakech, 2344–2351.

Zhang, Z., Iria, J., Brewster, Ch., Ciravegna, F. (2008). A Comparative Evaluation of Term Recognition Algorithms. In *Proceedings of LREC 2008* (CD-ROM). Marrakech, 2108–2113.

Zimányi, Á. (2006). *Nyelvhelyesség*. Eger, Eszterházy Károly Tanárképző Főiskola.

Mellékletek

1. PCT/FR2008/051836 szabadalom bibliográfiai adatai és leírása
2. PCT/FR2008/051836 terminusaira és terminusjelöltjeire vonatkozó statisztikai adatok

Nº de pub.:	WO/2009/053626	Numéro de la demande int.:	PCT/FR2008/051836
Date de la pub. int.:	30.04.2009	Date de dépôt int.:	09.10.2008

CIB: ***G06F 21/02*** (2006.01)

Déposants: **INGENICO FRANCE** [FR/FR]; 192, Avenue Charles de Gaulle F-92200 NEUILLY SUR SEINE (FR) *(Tous Sauf US)*.
ROLIN, Christian [FR/FR]; (FR) *(US Seulement)*.
TESTU, Dominique [FR/FR]; (FR) *(US Seulement)*.

Inventeurs: **ROLIN, Christian**; (FR).
TESTU, Dominique; (FR).

Mandataire: **CABINET BEAUMONT**; 1, Rue Champollion F-38000 GRENOBLE (FR) .

Données relatives à la priorité:
0758250 12.10.2007 FR

Titre: ECHANGE DE DONNEES ENTRE UN TERMINAL DE PAIEMENT ELECTRONIQUE ET UN OUTIL DE MAINTENANCE PAR UNE LIAISON USB

Abrégé: L’invention concerne un terminal de paiement électronique (30) comprenant une première borne de connexion USB (13) comportant au moins un premier fil (D+") de transfert de données. Le terminal de paiement comprend, en outre, une première résistance (24) reliant le premier fil à une première source (VDD) d’un premier potentiel et un interrupteur (38) entre le premier fil et la première résistance ou entre la première résistance et la première source.

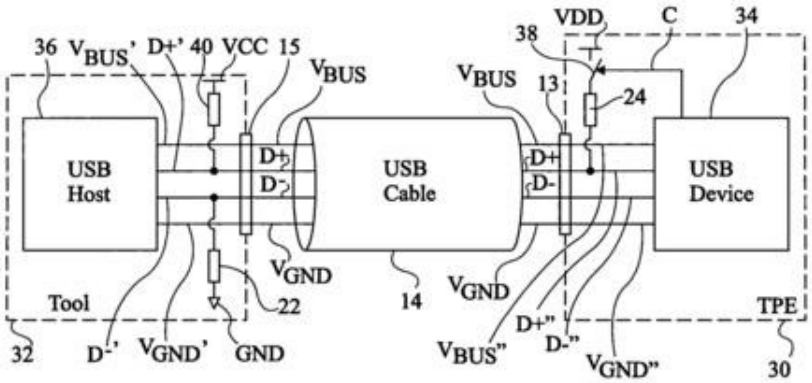


Fig 2

Leírás:

ECHANGE DE DONNEES ENTRE UN TERMINAL DE PAIEMENT ELECTRONIQUE ET UN OUTIL DE MAINTENANCE PAR UNE LIAISON USB

Domaine de l'invention

La présente invention concerne l'échange de données entre un terminal de paiement électronique et un outil de maintenance par l'intermédiaire d'une liaison USB (acronyme anglais pour Universal Serial Bus) .

Exposé de l'art antérieur

De plus en plus de terminaux de paiement électronique peuvent échanger des données avec un système externe, par exemple un ordinateur, par l'intermédiaire d'une liaison USB. A titre d'exemple, un terminal de paiement électronique peut être connecté à un ordinateur par l'intermédiaire d'une liaison USB afin que le terminal transmette à l'ordinateur des données, par exemple associées aux opérations de paiement réalisées par le terminal. Il est souhaitable de pouvoir utiliser la liaison USB du terminal de paiement électronique pour réaliser des opérations de maintenance.

Une opération de maintenance consiste, par exemple, à charger des logiciels dans une mémoire du terminal de paiement, à mettre à jour des logiciels déjà stockés dans le terminal de paiement, etc. Le niveau de sécurité généralement exigé dans le domaine des opérations de paiement impose qu'une opération de maintenance ne doit pas pouvoir être réalisée lors du fonctionnement normal du terminal de paiement électronique. Pour ce faire, une solution classique consiste à prévoir, au niveau du terminal, un interrupteur de sélection permettant de passer d'un mode de fonctionnement normal dans lequel une opération de maintenance est interdite à un mode de maintenance dans lequel une opération de maintenance peut être réalisée. De ce fait, si un ordinateur connecté au terminal de paiement tente d'effectuer une opération de maintenance alors que le terminal se trouve dans le mode de fonctionnement normal, le terminal refusera d'échanger des données avec l'ordinateur. Toutefois, le niveau de sécurité apporté par une telle solution peut s'avérer insuffisant.

En effet, l'interrupteur de sélection peut être actionné de façon frauduleuse pour sélectionner le mode de maintenance. Le terminal de paiement électronique peut alors être relié à un ordinateur classique par la liaison USB afin de réaliser une opération frauduleuse dans le mode de maintenance. Résumé de l'invention Un aspect de la présente invention vise un terminal de paiement électronique adapté à empêcher un échange de données avec un ordinateur classique par une liaison USB dans un mode de maintenance.

Selon un autre objet, le terminal de paiement électronique est adapté à réaliser au moins en partie automatiquement le passage entre le mode de fonctionnement normal et le mode de maintenance. Un autre aspect vise un système de maintenance qui, lorsque le terminal de paiement électronique est dans le mode de maintenance, est adapté à échanger des données avec le terminal par la liaison USB. Un autre aspect vise un procédé d'échange de données entre le terminal de paiement électronique et le système de maintenance par la liaison USB.

Dans ce but, elle prévoit un terminal de paiement électronique comprenant une première borne de connexion USB comportant au moins un premier fil de transfert de données. Le terminal de paiement comprend, en outre, une première résistance reliant le premier fil à une première source d'un premier potentiel et un interrupteur entre le premier fil et la première résistance ou entre la première résistance et la première source. Selon un exemple de réalisation, le terminal comprend un circuit de commande adapté à fermer l'interrupteur

dans un premier mode de fonctionnement et à ouvrir l'interrupteur dans un second mode de fonctionnement. Le terminal est destiné à être relié, par la première borne de connexion USB, à un premier système dans le premier mode de fonctionnement et à un second système dans le second mode de fonctionnement. Les premier et second systèmes sont adaptés à initier un échange de données avec le terminal. Le premier système comprend un deuxième fil destiné à être relié au premier fil et une deuxième résistance reliant le deuxième fil à une deuxième source d'un deuxième potentiel strictement inférieur au premier potentiel. Le second système comprend un troisième fil destiné à être relié au premier fil et une troisième résistance reliant le troisième fil à une troisième source d'un troisième potentiel strictement supérieur au deuxième potentiel. II est également prévu un système comprenant au moins une seconde borne de connexion USB comportant au moins un quatrième fil de transfert de données. Le système comprend un circuit adapté à transmettre, par la seconde borne de connexion USB, de façon répétée une requête d'identification pour débuter un échange de données par la seconde borne de connexion tant qu'un signal de réponse n'est pas reçu depuis la seconde borne de connexion. Selon un exemple de réalisation, le système est destiné à être relié à un terminal tel que défini précédemment.

Selon un exemple de réalisation, le système comprend une quatrième source d'un quatrième potentiel, une cinquième source d'un cinquième potentiel strictement supérieur au quatrième potentiel et une quatrième résistance reliant le quatrième fil à la cinquième source. Il est également prévu un procédé d'échange de données entre un terminal de paiement électronique et un premier système ou un second système . Le terminal comprend une borne de connexion USB comprenant au moins un premier fil de transfert de données. Le terminal comprend, en outre, une première résistance reliant le premier fil à une première source d'un premier potentiel, et un interrupteur entre le premier fil et la première résistance ou entre la première résistance et la première source. Les premier et second systèmes sont adaptés à initier un échange de données avec le terminal. Le premier système comprend un deuxième fil destiné à être relié au premier fil et une deuxième résistance reliant le deuxième fil à une deuxième source d'un deuxième potentiel strictement inférieur au premier potentiel . Le second système comprend un troisième fil destiné à être relié au premier fil et une troisième résistance reliant le troisième fil à une troisième source d'un troisième potentiel strictement supérieur au deuxième potentiel. Le procédé comprend les étapes consistant à amener le terminal à ouvrir l'interrupteur pour permettre l'échange de données entre le terminal et le second système ; et à amener le terminal à fermer l'interrupteur pour permettre l'échange de données entre le terminal et le premier système . Selon un exemple de réalisation, le procédé consiste à amener le terminal, à la mise sous tension du terminal, à ouvrir l'interrupteur pendant une durée donnée, et, à fermer ensuite l'interrupteur si, pendant ladite durée, un échange de données n'a pas eu lieu entre le terminal et le second système. Selon un exemple de réalisation, le procédé consiste à amener le terminal à ne pas répondre à des requêtes fournies par le second système lorsque l'interrupteur est fermé.

Selon un exemple de réalisation, le procédé consiste à amener le second système à transmettre, au moins en partie par le troisième fil, de façon répétée une requête d'identification pour débuter un échange de données avec le terminal tant que le second système ne reçoit pas de réponse du terminal. Brève description des dessins Ces objets, caractéristiques et avantages, ainsi que d'autres seront exposés en détail dans la description suivante d'un exemple de réalisation particulier faite à titre non-limitatif en relation avec les figures jointes parmi lesquelles : la figure 1 représente, de façon schématique, un exemple classique de liaison entre un terminal de paiement électronique et un ordinateur par un câble USB ; la figure 2 représente, de façon schématique, un exemple de liaison

selon l'invention entre un terminal de paiement électronique et un outil de maintenance par un câble USB ; et la figure 3 représente, sous la forme d'un schéma par blocs, un exemple de procédé de fonctionnement du terminal. Description détaillée Par souci de clarté, de mêmes éléments ont été désignés par de mêmes références aux différentes figures. Seuls les éléments nécessaires à la compréhension de la présente invention sont représentés sur les figures et seront décrits par la suite. La figure 1 représente, de façon classique, un terminai de paiement électronique 10 (TPE) relié à un ordinateur 12 (PC) . Le terminal 10 comprend une borne de connexion USB 13 dans laquelle est connectée une extrémité d'un câble USB 14 (USB Cable) . L'ordinateur 12 comprend une borne de connexion USB 15 recevant l'extrémité opposée du câble USB 14. De façon classique, le terminal de paiement électronique 10 permet de réaliser des opérations de paiement, par exemple par l'intermédiaire d'une carte à puce, d'une carte magnétique, d'un chèque, etc... Comme cela est décrit dans la norme USB 2.0, le câble USB 14 comprend quatre fils conducteurs ou lignes conductrices.

Le fil conducteur Vgyg sert à la transmission d'un potentiel de référence haut, généralement de quelques volts. Le fil conducteur V GND sert à la transmission d'un potentiel de référence bas, généralement la masse de l'ordinateur 12. Les fils conducteurs D+ et D- servent à la transmission du signal utile. L'ordinateur 12 comprend un module de communication 16 (USB Host) , appelé module hôte 16 dans la suite de la description. La borne de connexion 15 comprend quatre fils Vgyg ' , D+', D-' et V GND* prolongent respectivement les fils Vgyg, D+, D- et V GND jusqu'au module hôte 16. Le terminal 10 comprend un module de communication 18 (USB Device) , appelé module périphérique dans la suite de la description. La borne de connexion 13 comprend quatre fils Vgijg", D+", D-" et V Q Q" qui prolongent respectivement les fils Vgyg, D+, D- et V Q ^ Q jusqu'au module périphérique 18. Les modules 16 et 18 sont adaptés à échanger des données par l'intermédiaire du câble USB 14, par exemple selon le protocole d'échange de données décrit dans la norme USB 2.0. En particulier, l'échange de données est initié par le module hôte 16. De façon classique, on prévoit au niveau de l'ordinateur 12 des résistances 20, 22 (généralement appelées résistances Pull Down) reliant les fils D+ ' et D-' à la masse GND. Du côté du terminal 10, et, de façon générale, du côté de n'importe quel appareil périphérique adapté à être connecté à un ordinateur par une liaison USB, il est prévu une résistance 24 (généralement appelée résistance Pull Up) qui relie le fil D+" (ou, à titre de variante, le fil D-") à une source VDD d'un potentiel haut généralement de quelques volts. La résistance Pull Up 24 est plus faible, généralement d'un ordre de grandeur, que les résistances Pull Down 20, 22. La source VDD peut correspondre au fil Vg]Jg". Lorsque le terminal 10 n'est pas connecté à l'ordinateur 12, les fils D+, D+', D- et D-' sont sensiblement maintenus au potentiel de la masse GND par l'intermédiaire des résistances 20, 22 côté ordinateur 12. Lorsque le terminal 10

est connecté à l'ordinateur 12 par le câble USB 14, le module hôte 16 détecte la présence du terminal 10 par l'élévation du potentiel des fils D+', D+, D+" due à la résistance 24 côté terminal 10 qui relie le fil D+" à la source VDD. Le module hôte 16 initie alors un échange de données avec le terminal 10, généralement par l'envoi d'une requête d'identification au terminal 10 pour obtenir les paramètres de fonctionnement du terminal (par exemple la requête Getdescriptor définie par la norme USB 2.0) . En l'absence de réponse du terminal 10, le module hôte 16 transmet à nouveau la requête d'identification à deux reprises. Si le terminal 10 ne répond toujours pas, le module hôte 16 considère que le module périphérique 18 est hors service et la communication est interrompue. Le terminal 10 comprend un module de sélection 25 (SW) adapté à fournir un signal S de sélection de modes de fonctionnement au module périphérique 18. Le module de sélection 25 peut être un interrupteur mécanique. Selon la valeur du signal S transmis par le module de sélection

25, le module périphérique 18 fonctionne dans un mode de fonctionnement normal ou dans un mode de maintenance. Dans chacun de ces modes, le module périphérique 18 est adapté à échanger des données sur la liaison USB 14 selon un protocole particulier et s'attend notamment à recevoir des requêtes d'un type particulier. Pour éviter qu'une opération de maintenance puisse être réalisée par un ordinateur classique pour permettre que le passage entre le mode de fonctionnement normal et le mode de maintenance soit effectué, au moins en partie, de façon automatique par le terminal, il est prévu de modifier la structure du terminal de sorte qu'une opération de maintenance ne puisse être réalisée que par un outil de maintenance spécifique ayant une structure différente de celle d'un ordinateur classique. En outre, le protocole d'échange entre l'outil de maintenance et le terminal est modifié par rapport au protocole d'échange classique décrit dans la norme USB 2.0.
Plus précisément, on prévoit que, dans le mode de maintenance, le terminal est adapté à temporairement "retirer" la résistance Pull Up de façon à ne pas élever le potentiel des fils D+', D+, D+" lorsque le terminal est relié à l'outil de maintenance (ou à un ordinateur) par une liaison USB. En outre, il est prévu que l'outil de maintenance envoie en permanence une requête d'identification jusqu'à ce qu'il reçoive une réponse de la part du terminal. De ce fait, c'est le terminal qui détecte la présence de l'outil de maintenance par la réception d'une requête d'identification et non l'outil de maintenance qui détecte la présence du terminal par une élévation du potentiel des fils D+', D+, D+". En mode de fonctionnement normal, le terminal est adapté à "remettre en place" la résistance Pull Up pour permettre un échange de données classique selon la norme USB classique. La figure 2 représente un exemple de réalisation du terminal de paiement électronique 30 (TPE) et de l'outil de maintenance 32 (Tool) . Les éléments communs avec le terminal 10 et l'ordinateur 12 représentés en figure 1 sont désignés par les mêmes références. Le terminal 30 comprend un module de communication 34 (USB Device) , appelé module périphérique, fonctionnant de façon analogue au module 18 sauf sur certains points qui vont être décrits par la suite.
De même, l'outil de maintenance 32 (Tool) comprend un module de communication 36 (USB Host) , appelé module hôte, fonctionnant de façon analogue au module 16 à la différence que le protocole de communication mis en oeuvre par le module 36 diffère en certains points du protocole de communication classique mis en oeuvre par le module 16. Le terminal 30 comprend un interrupteur 38 disposé, par exemple, entre la résistance 24 et la source de potentiel de référence VDD. L'interrupteur 38 est commandé par un signal C, fourni par le module périphérique 34. L'outil de maintenance 32 comprend une résistance 40 du type Pull Up reliant le fil D+ ' à une source d'un potentiel de référence haut VCC. La source VCC peut correspondre au fil Vgyg ' . L'outil 32 comprend, en outre, la résistance Pull Down 22 reliant le fil D-' à la masse GND, la résistance 20 reliant le fil D+ ' à la masse n'étant pas présente. Les résistances 40, 22 assurent une polarisation convenable des fils D+ ' et D-' pour permettre la transmission de données sur la liaison USB 14. Dans le mode de fonctionnement normal, le module périphérique 34 commande la fermeture de l'interrupteur 38 de sorte que le terminal 30 puisse fonctionner de façon analogue au terminal 10 représenté en figure 1.
En particulier, le terminal 30 peut alors être relié de façon classique à un ordinateur par la liaison USB 14. Dans le mode de maintenance, le module périphérique 34 commande l'ouverture de l'interrupteur 38. Dans ce cas, si le terminal 30 est connecté à un ordinateur classique, tel que l'ordinateur 12 de la figure 1, l'ordinateur 12 ne peut pas détecter la présence du terminal 30 puisque les fils D+ ' et D-' restent à la masse. Aucun échange de données ne peut alors se produire. La figure 3 illustre les étapes d'un exemple de procédé de fonctionnement du terminal 30. A l'étape 50, le terminal 30 est mis sous tension. Le module périphérique 34 commande l'ouverture de l'interrupteur 38. Le terminal 30 est alors dans le mode de maintenance. Le procédé se poursuit à l'étape 52. A l'étape 52, le

terminal 30 attend la réception d'une requête de l'outil de maintenance 32. Si, au bout d'une durée déterminée, aucun échange de données n'a eu lieu entre le terminal 30 et l'outil de maintenance 32, le procédé se poursuit à l'étape 54. A l'étape 54, le module périphérique 34 commande la fermeture de l'interrupteur 38. Le terminal 30 est alors dans le mode de fonctionnement normal. Une fois que le terminal 30 est dans le mode de fonctionnement normal, il est nécessaire, pour réaliser une opération de maintenance, d'éteindre puis de remettre sous tension le terminal 30.

A l'étape 52, si le terminal 30 reçoit une requête de l'outil de maintenance 32, le procédé se poursuit à l'étape 56. A l'étape 56, une opération de maintenance a lieu. Lorsqu'elle est achevée, le procédé se poursuit à l'étape 54 dans laquelle le terminal 30 passe en mode de fonctionnement normal . A l'étape 52, lorsque le terminal 30 est relié à l'outil de maintenance 32 par la liaison USB 14, l'outil de maintenance 32 n'est pas en mesure de détecter la présence du terminal 30 puisque le fil D+' est à un potentiel haut par l'intermédiaire de la résistance pull up 40 indépendamment de la présence ou de l'absence du terminal 30. De ce fait, si on utilisait le protocole USB d'échanges de données classique, le module hôte 36 croirait détecter la présence du terminal 30 dès sa mise en marche et transmettrait immédiatement trois fois la requête d'identification et, du fait qu'il serait fort peu probable que le terminal 30 soit présent à cet instant, le module hôte 36 empêcherait tout échange de données ultérieur. Pour qu'un échange de données puisse être effectué, il est nécessaire de modifier les étapes initiales du protocole USB mis en oeuvre par le module hôte 36. Pour ce faire, on prévoit que le module hôte 36 transmette sans arrêt une requête d'identification jusqu'à ce qu'il reçoive une réponse de la part du terminal 30. De ce fait, lorsque le terminal 30 est dans le mode de maintenance, le module périphérique 34 détecte la présence de l'outil de maintenance 30 par la réception d'une requête d'identification transmise par les fils D+ et D-. Le module périphérique 34 peut alors répondre au module hôte 36 et l'échange de données peut se poursuivre de façon classique à l'étape 56. Si le terminal 30 est dans le mode de fonctionnement normal, il suffit qu'il ne réponde pas à la requête d'identification du module hôte 36 pour empêcher qu'une opération de maintenance ait lieu. Selon une variante, l'outil de maintenance 32 peut être autonome ou connecté lui-même à un ordinateur.

Des modes de réalisation particuliers de la présente invention ont été décrits. Diverses variantes et modifications apparaîtront à l'homme de l'art. En particulier, bien que la présente invention ait été décrite pour la maintenance d'un terminal de paiement électronique, il est clair que la présente invention peut s'appliquer à tout type d'opérations réalisées auprès d'un terminal de paiement électronique pour lesquelles on souhaite imposer l'utilisation d'un outil spécifique et non d'un ordinateur classique. Il s'agit, par exemple, d'une opération de diagnostic.

	Terminus	Tanulókorpuszban	Szűrés nélküli listában	Szűrt listában	Helyesen kinyert	A szabály alapú kinyerés és szűrés hibaforrásai	Legnagyobb F-érték	Legnagyobb pontosság	Előfordulás	Webes előfordulás	Weirdness	C-érték	NC-érték	ÖÉ
1	absence		+	+		nem terminus	+		1	16774	0,0121	3,40E+038	1,5974	0,1291
2	absence de réponse		+	+		a prepozíció és az az előtti rész nem része a terminusnak	+		1	1136	0,1789	1,5850	1,5974	0,5107
3	acronyme anglais		+	+		nem terminus	+		1	12	16,9328	1	1,4492	0,5594
4	appareil périphérique	+		+	+		+		1	0	3,40E+038	1	1	0,2
5	appareil périphérique adapté		+											
6	arrêt		+	+		nem terminus			1	67737	0,0030	3,40E+038	1,0598	0,0144
7	art	+	+	+	+		+		2	256083	0,0016	3,40E+038	1,5149	0,1043
8	art antérieur		+											
9	aspect		+	+		nem terminus	+		3	4269	0,1428	3,40E+038	1,5289	0,2122
10	avantage		+	+		nem terminus	+		1	32456	0,0063	3,40E+038	1,4566	0,0963
11	borne de connexion	+	+	+	+		+		4	2	406,3861	-4,7549	1,5974	0,6780

12	borne de connexion usb	+	+	+	+		+		7	0	3,40E+038	14	1,4486	0,5588
13	bout		+	+			+		1	244778	8,30E-004	3,40E+038	1,5149	0,1036
14	but		+											
15	câble usb	+	+	+	+		+		7	28	50,7983	7	1,5289	0,6231
16	carte à puce	+	+	+	+		+		1	130	1,5630	1,5850	1,4486	0,5169
17	carte magnétique	+	+	+	+		+		1	22	9,2360	1	1,4486	0,5588
18	cas		+											
19	chèque	+	+	+	+		+		1	5525	0,0368	3,40E+038	1,5289	0,1347
20	circuit	+		+	+		+		1	14100	0,0144	3,40E+038	1,5289	0,1172
21	circuit adapté		+											
22	circuit de commande	+		+	+		+		1	230	0,8834	1,5850	1,5289	0,5405
23	circuit de commande adapté		+											
24	clarté			+		nem terminus	+		1	1778	0,1143	3,40E+038	1,4751	0,1814
25	communication	+	+	+	+		+		1	52383	0,0039	3,40E+038	1,5974	0,1226
26	compréhension		+	+		nem terminus	+		1	3546	0,0573	3,40E+038	1,5974	0,1640
27	côté		+			nem terminus								
28	côté ordinateur		+	+		nem terminus	+		1	1	203,1930	1	1,1833	0,3466
29	côté terminal		+	+		nem terminus	+		1	0	3,40E+038	1	1,1209	0,2967
30	d		+	+		a jel része lenne a terminusnak	+		22	516948	0,0086	3,40E+038	1,7748	0,1618
31	d -	+				a jel része lenne a terminusnak								

32	d +	+				a jel része lenne a terminusnak								
33	description	+	+	+	+		+		5	19243	0,0528	3,40E+038	1,5974	0,1606
34	description suivant		+											
35	dessin	+	+	+	+		+		1	25746	0,0079	3,40E+038	1,5512	0,1165
36	détail		+	+		nem terminus			1	59179	0,0034	3,40E+038	1,0448	0,0117
37	différence		+	+		nem terminus	+		1	21014	0,0097	3,40E+038	1,5974	0,1272
38	domaine	+	+	+	+		+		2	32030	0,0127	3,40E+038	1,5149	0,1131
39	donnée	+	+	+	+		+		5	22604	0,0449	3,40E+038	1,5512	0,1454
40	durée		+	+		nem terminus	+		1	28554	0,0071	3,40E+038	1,3301	0,0717
41	durée déterminé		+	+		nem terminus	+		1	191	1,0638	1	1,4486	0,4898
42	durée donné		+	+		nem terminus	+		1	6	33,8655	1	1,4486	0,5588
43	échange		+	+		a terminus nagyobb	+		1	31958	0,0064	3,40E+038	1,5289	0,1109
44	échange de donnée	+	+	+	+		+		16	257	12,6501	23,3782	1,5289	0,6231
45	échange de donnée classique		+	+		a melléknév nem része a terminusnak	+		1	13	15,6302	0	1,5289	0,6231
46	échange de donnée ultérieur		+											
47	effet		+											
48	élément		+	+		nem terminus	+		1	22251	0,0091	3,40E+038	1,2392	0,0551

49	élément commun		+	+		nem terminus	+		1	30	6,7731	1	1,5512	0,6408
50	élément nécessaire		+	+		nem terminus	+		1	29	7,0067	1	1,5512	0,6408
51	élévation	+	+	+	+		+		2	700	0,5806	3,40E+038	1,5974	0,4718
52	envoi		+	+		rossz az automatikus szófaji címke/lemma	+		1	15168	0,0134	3,40E+038	1,5149	0,1136
53	est également prévu		+											
54	est prévu			+		rossz az automatikus szófaji címke/lemma			1	4700	0,0432	1	1	0,0085
55	étape	+	+	+	+		+		13	37466	0,0705	3,40E+038	1,5974	0,1739
56	étape initial	+	+	+	+		+		1	1	203,1930	1	1,5512	0,6410
57	exemple		+											
58	exemple classique de liaison		+	+		nem terminus	+		1	0	340282350000000000000000000000000000,0000	2	1,5289	0,6231
59	exemple de liaison		+	+		nem terminus	+		1	42	4,8379	1,5850	1,5289	0,6216
60	exemple de procédé de fonctionnement	+	+	+	+		+		2	20	20,3193	4,6439	1,5289	0,6231
61	exemple de réalisation	+	+	+	+		+		7	87	16,3489	9,5098	1,5289	0,6231

62	exemple de réalisation particulier	+	+	+	+		+		1	36	5,6443	2	1,5289	0,6224
63	extrêmité	+	+	+	+		+		1	866	0,2346	3,40E+038	1,4486	0,2570
64	extrêmité opposé	+	+	+	+		+		1	1	203,1930	1	1,5974	0,6780
65	façon analogue		+											
66	façon automatique		+											
67	façon classique		+											
68	façon frauduleux		+											
69	façon général		+											
70	façon répété		+											
71	façon schématique		+											
72	fait		+											
73	fermeture	+	+	+	+		+		2	8120	0,0500	3,40E+038	1,5974	0,1585
74	figure	+	+	+	+		+		12	43700	0,0558	3,40E+038	1,5974	0,1629
75	fil	+	+	+	+		+		20	136560	0,0298	3,40E+038	1,5149	0,1264
76	fil conducteur	+	+	+	+		+		2	93	4,3697	2	1,5149	0,6094
77	fil conducteur d -	+				a jel része lenne a terminusnak								
78	fil conducteur d +	+				a jel része lenne a terminusnak								

79	fil d		+	+		a jel része lenne a terminusnak	+		7	4235	0,3359	4	1,5149	0,4690
80	fil d -	+				a jel része lenne a terminusnak								
81	fil d +	+				a jel része lenne a terminusnak								
82	fil de transfert de donnée	+	+	+	+		+		3	16	38,0987	6,9658	1,1718	0,3375
83	fils		+	+		rossz az automatikus szófaji címke/lemma	+		3	20315	0,0300	3,40E+038	1,5512	0,1339
84	fils conducteur		+	+		rossz az automatikus szófaji címke/lemma	+		1	14	14,5138	0	1	0,2000
85	fils conducteur d		+	+		rossz az automatikus szófaji címke/lemma	+		1	11	18,4721	1,5850	1,5512	0,6410
86	fils d		+	+		rossz az automatikus szófaji címke/lemma	+		8	1095	1,4845	8	1,5512	0,5957
87	fils vgijg		+	+		rossz az automatikus szófaji címke/lemma	+		1	0	3,40E+038	1	1	0,2
88	fois		+											
89	fonctionnement normal	+	+	+	+		+		1	91	2,2329	-10	1,5149	0,5905
90	forme		+											
91	gnd		+	+		a jel része lenne a terminusnak	+		2	4	101,5965	3,40E+038	1	0,8
92	gnd *	+				a jel része lenne a terminusnak								
93	homme		+	+		a terminusban van determináns	+		1	147035	0,0014	3,40E+038	1,5149	0,1041
94	homme de le art	+				a terminusban van determináns								
95	hôte		+	+		terminus nagyobb			2	33568	0,0121	3,40E+038	1	0,0096

96	instant		+	+		nem terminus			1	22858	0,0089	3,40E+038	1,1625	0,0396
97	intermédiaire		+	+		nem terminus	+		7	5171	0,2751	3,40E+038	1,5974	0,3119
98	interrupteur	+	+	+	+		+	+	15	120	25,3991	3,40E+038	1,5289	0,9058
99	interrupteur de sélection	+	+	+	+		+		2	1	406,3861	3,1699	1,5289	0,6231
100	interrupteur mécanique	+	+	+	+		+		1	0	3,40E+038	1	1,5289	0,6231
101	invention	+	+	+	+		+		9	3286	0,5565	3,40E+038	1,8987	0,5212
102	jour		+	+		nem terminus			1	369283	5,50E-004	3,40E+038	1,2390	0,0482
103	liaison usb	+	+	+	+		+		15	8	380,9869	15	1,5974	0,6780
104	lieu		+	+		nem terminus			4	160548	0,0051	3,40E+038	1,1402	0,0321
105	ligne conducteur	+	+	+	+		+		1	1	203,1930	1	1,1850	0,3480
106	logiciel	+	+	+	+		+		2	17026	0,0239	3,40E+038	1,5512	0,1291
107	maintenance	+	+	+	+		+		1	2357	0,0862	3,40E+038	1,5974	0,1856
108	masse	+	+	+	+		+		3	16608	0,0367	3,40E+038	1,5974	0,1483
109	masse gnd	+	+	+	+		+		3	0	3,40E+038	3	1,5974	0,6780
110	mémoire	+	+	+	+		+		1	29438	0,0069	3,40E+038	1,4486	0,0952
111	mesure de détecter		+											
112	mise en marche	+	+	+	+		+		1	7202	0,0282	1,5850	1,1008	0,0862
113	mise sous tension	+	+	+	+		+		1	54	3,7628	1,5850	1,5974	0,6733
114	mode	+	+	+	+		+		1	290583	6,99E-004	3,40E+038	1,2798	0,0565
115	mode à ouvrir		+	+			+		1	115	1,7669	1,5850	1,1718	0,3033

116	mode de fonctionnement	+	+	+	+		+		4	795	1,0224	-3,1699	1,2502	0,3282
117	mode de fonctionnement normal	+	+	+	+		+		11	64	34,9238	22	1,5289	0,6231
118	mode de maintenance	+	+	+	+		+		12	68	35,8576	19,01955	1,5289	0,6231
119	mode de réalisation particulier	+	+	+	+		+		1	95	2,1389	2	1,5064	0,5816
120	modification		+	+		nem terminus	+		1	8740	0,0232	3,40E+038	1,4566	0,1097
121	module	+	+	+	+		+		5	1749	0,5809	3,40E+038	1,5512	0,4627
122	module de communication	+	+	+	+		+		4	61	13,3241	6,3399	1,5289	0,6231
123	module de sélection	+	+	+	+		+		3	77	7,9166	4,7549	1,5289	0,6231
124	module hôte	+	+	+	+		+		12	0	3,40E+038	12	1,5149	0,6119
125	module périphérique	+	+	+	+		+		14	0	3,40E+038	14	1,5149	0,6119
126	moins		+											
127	moins en partie		+											
128	nd		+	+		beolvasási hiba			1	14716	0,0138	3,40E+038	1	0,0110
129	niveau		+	+		nem terminus	+		2	90291	0,0045	3,40E+038	1,5149	0,1066
130	niveau de sécurité	+	+	+	+		+		2	1429	0,2844	3,1699	1,5149	0,4614
131	norme usb	+	+	+	+		+		4	2	406,3861	3	1,5974	0,6780

132	norme usb classique		+	+		a melléknév nem része a terminusnak	+		1	0	3,40E+038	1,5850	1,5974	0,6780
133	objet		+	+		nem terminus	+		2	44290	0,0092	3,40E+038	1,2798	0,0633
134	oeuvre		+											
135	opération de diagnostic	+	+	+	+		+		1	72	2,8221	1,5850	1,4486	0,5470
136	opération de maintenance	+	+	+	+		+		11	134	16,6800	17,4346	1,5064	0,6051
137	opération de paiement	+	+	+	+		+		3	195	3,1260	4,7549	1,5512	0,6322
138	opération frauduleux	+	+	+	+		+		1	0	3,40E+038	1	1,4486	0,5588
139	ordinateur	+	+	+	+		+		20	8355	0,4864	3,40E+038	1,5289	0,4139
140	ordinateur classique	+	+	+	+		+		6	5	243,8316	6	1,5289	0,6231
141	ordre de grandeur	+	+	+	+		+		1	334	0,6084	1,5850	1,5289	0,5143
142	outil	+	+	+	+		+		1	84965	0,0024	3,40E+038	1,5149	0,1049
143	outil de maintenance	+	+	+	+		+		18	27	135,4620	26,9444	1,5289	0,6231
144	outil de maintenance spécifique		+	+		a melléknév nem része a terminusnak	+		1	3	67,7310	2	1,5289	0,6231
145	outil spécifique		+	+		a melléknév nem része a terminusnak	+		1	13	15,6302	1	1,5289	0,6231
146	ouverture	+	+	+	+		+		2	51537	0,0079	3,40E+038	1,5974	0,1258

147	paiement		+	+		rossz az automatikus szófaji címke/lemma	+		4	8600	0,0945	3,40E+038	1,8146	0,2351
148	paiement électronique		+	+		rossz az automatikus szófaji címke/lemma	+		19	78	49,4957	18	1,8146	0,8517
149	paiement électronique adapté		+											
150	paiement électronique comprenant		+	+		rossz az automatikus szófaji címke/lemma	+		1	0	3,40E+038	1,5850	1,8146	0,8517
151	paramètre de fonctionnement	+	+	+	+		+		1	7	29,0276	1,5850	1,5512	0,6410
152	part		+	+		nem terminus	+		2	390099	0,0010	3,40E+038	1,5974	0,1203
153	particulier		+											
154	partie			+		nem terminus			2	106146	0,0038	3,40E+038	1,0448	0,0120
155	passage	+	+	+	+		+		2	56496	0,0072	3,40E+038	1,5149	0,1087
156	pc	+	+	+	+		+		1	6496	0,0313	3,40E+038	1,4492	0,1145
157	permanence		+	+		nem terminus			1	5394	0,0377	3,40E+038	1,0448	0,0385
158	place		+	+		nem terminus			1	190070	0,0011	3,40E+038	1,0448	0,0098
159	point		+	+		nem terminus			2	130451	0,0031	3,40E+038	1	0,0025
160	polarisation	+				terminus rövidebb								
161	polarisation convenable		+	+		a melléknév nem része a terminusnak	+		1	0	3,40E+038	1	1,4486	0,5588
162	potentiel	+	+	+	+		+		12	19104	0,1276	3,40E+038	1,5149	0,1988
163	potentiel de référence		+	+		rossz az automatikus szófaji címke/lemma	+		2	170	2,3905	1,5849625	1,5289	0,6048

164	potentiel de référence bas	+				rossz az automatikus szófaji címke/lemma								
165	potentiel de référence haut	+				rossz az automatikus szófaji címke/lemma								
166	potentiel de référence haut vcc	+	+	+	+		+		1	0	3,40E+038	2,3219	1,5289	0,6231
167	potentiel haut	+	+	+	+		+		2	1	406,3861	2	1,5289	0,6231
168	potentiel strictement inférieur		+	+		a melléknév nem része a terminusnak	+		2	0	3,40E+038	3,1699	1,1328	0,3063
169	potentiel strictement supérieur		+	+		a melléknév nem része a terminusnak	+		3	0	3,40E+038	4,7549	1,2370	0,3896
170	présence		+	+		nem terminus	+		8	32959	0,0493	3,40E+038	1,5974	0,1580
171	procédé	+	+	+	+		+		8	15654	0,1038	3,40E+038	1,5149	0,1819
172	procédé de échange de donnée	+	+	+	+		+		2	5	81,2772	4,6439	1,5289	0,6231
173	procédé de fonctionnement	+				terminus rövidebb								
174	protocole	+				a melléknév nem része a terminusnak								
175	protocole de communication	+	+	+	+		+		1	108	1,8814	0	1,5149	0,5815

176	protocole de communication classique	+	+	+	+		+		1	7	29,0276	2	1,5149	0,6119
177	protocole de échange	+	+	+	+		+		1	67	3,0327	0	1,5149	0,6023
178	protocole de échange classique			+		a melléknév nem része a terminusnak	+		1	8	25,3991	2	1,5149	0,6119
179	protocole de échange classique décrit		+											
180	protocole de échange de donnée	+	+	+	+		+		1	21	9,6759	2,3219	1,5149	0,6119
181	protocole particulier		+	+		a melléknév nem része a terminusnak	+		1	1	203,1930	1	1,5289	0,6231
182	protocole usb	+	+	+	+		+		1	0	3,40E+038	0	1,5149	0,6119
183	protocole usb de échange de donnée classique	+	+	+	+		+		1	0	3,40E+038	2,8074	1,5149	0,6119
184	pull up	+	+	+	+		+		1	0	3,40E+038	-1,5	1,0781	0,2624
185	réception	+	+	+	+		+		3	5887	0,1035	3,40E+038	1,5974	0,1982
186	référence	+	+	+	+		+		2	35931	0,0113	3,40E+038	1,2392	0,0568
187	réponse	+	+	+	+		+		3	62416	0,0098	3,40E+038	1,8146	0,1707
188	reprise		+	+		nem terminus			1	50297	0,0040	3,40E+038	1,1173	0,0267
189	requête	+	+	+	+		+		4	2332	0,3485	3,40E+038	1,5512	0,3457

190	requête de identification	+	+	+	+		+		10	1	2031,9303	15,8496	1,5974	0,6780
191	requête getdescriptor	+	+	+	+		+		1	0	3,40E+038	1	1,5974	0,6780
192	résistance	+	+	+	+		+		20	25347	0,1603	3,40E+038	1,5974	0,2380
193	résistance pull down	+	+	+	+		+		3	0	3,40E+038	4,7549	1,5974	0,6780
194	résistance pull up	+	+	+	+		+		4	0	3,40E+038	6,3399	1,5974	0,6780
195	schéma par bloc	+	+	+	+		+		1	0	3,40E+038	1,5850	1,5289	0,6231
196	second système		+	+		rossz az automatikus szófaji címke/lemma	+		2	8	50,7983	2	1,4566	0,5653
197	service		+	+		nem terminus			1	206272	9,85E-004	3,40E+038	1	7,88E-004
198	signal	+												
199	signal c	+	+	+	+		+		1	9	22,5770	1	1,5289	0,6231
200	signal de réponse	+	+	+	+		+		1	64	3,1749	1,5850	1,5289	0,6148
201	signal de sélection de mode de fonctionnement	+												
202	signal s		+	+			+		1	11	18,4721	0	1,5149	0,6119

203	signal s de sélection de mode de fonctionnement		+	+			+		1	0	3,40E+038	3	1,5289	0,6231
204	signal utile	+	+	+	+		+		1	0	3,40E+038	1	1,5149	0,6119
205	solution		+	+		nem terminus	+		1	102228	0,0020	3,40E+038	1,4746	0,0965
206	solution classique		+	+		a melléknév nem része a terminusnak	+		1	5	40,6386	1	1,4486	0,5588
207	souci de clarté		+											
208	source	+	+	+	+		+		12	117677	0,0207	3,40E+038	1,4486	0,1061
209	source de potentiel de référence vdd	+	+	+	+		+		1	0	3,40E+038	2,5850	1,5974	0,6780
210	source vcc	+	+	+	+		+		1	0	3,40E+038	1	1,5974	0,6780
211	source vdd	+	+	+	+		+		3	0	3,40E+038	3	1,5974	0,6780
212	structure	+	+	+	+		+		2	12020	0,0338	3,40E+038	1,5974	0,1461
213	structure différent		+											
214	suite		+	+		nem terminus	+		4	65461	0,0124	3,40E+038	1,5974	0,1294
215	sw	+	+	+	+		+		1	242	0,8396	3,40E+038	1,4492	0,5444
216	système	+	+	+	+		+		17	51943	0,0665	3,40E+038	1,5289	0,1573
217	système à transmettre		+	+			+		1	21	9,6759	1,5850	1,2502	0,4001
218	système de maintenance	+	+	+	+		+		2	156	2,6050	3,1699	1,5289	0,6084

219	système externe	+	+	+	+		+		1	7	29,0276	1	1,5289	0,6231
220	tension	+	+	+	+		+		2	21513	0,0189	3,40E+038	1,1673	0,0484
221	terminal	+				rossz az automatikus szófaji címke/lemma								
222	terminal de paiement	+				rossz az automatikus szófaji címke/lemma								
223	terminal de paiement électronique	+				rossz az automatikus szófaji címke/lemma								
224	titre de exemple		+											
225	titre de variante		+	+		nem terminus	+		1	29	7,0067	1,5850	1,2390	0,3910
226	titre non-limitatif en relation		+											
227	tool	+	+	+	+		+		1	1265	0,1606	3,40E+038	1,4492	0,2086
228	tpe	+	+	+	+		+		2	5308	0,0766	3,40E+038	1,4492	0,1488
229	transmission	+	+	+	+		+		3	7070	0,0862	3,40E+038	1,5974	0,1856
230	transmission de donnée	+	+	+	+		+		1	134	1,5164	1,5850	1,5974	0,6341
231	type	+				a prepozíció és az az előtti rész nem része a terminusnak								
232	type de opération		+	+		a prepozíció és az az előtti rész nem része a terminusnak	+		1	1328	0,1530	1,5850	1,1007	0,1089
233	type particulier		+	+		a prepozíció és az az előtti rész nem része a terminusnak	+		1	9	22,5770	1	1,5289	0,6231

234	type pull up		+	+		a prepozíció és az az előtti rész nem része a terminusnak	+		1	0	3,40E+038	1,5850	1,5149	0,6119
235	universal serial bus	+	+	+	+		+		1	0	3,40E+038	1,5850	1,2411	0,3929
236	usb câble	+	+	+	+		+		1	0	3,40E+038	1	1,4492	0,5594
237	usb dévice	+	+	+	+		+		2	0	3,40E+038	2	1,4492	0,5594
238	usb host	+	+	+	+		+		2	0	3,40E+038	2	1,4492	0,5594
239	utilisation	+	+	+	+		+		1	14266	0,0142	3,40E+038	1,5974	0,1308
240	valeur	+	+	+	+		+		1	76777	0,0026	3,40E+038	1,5974	0,1216
241	variante		+	+		nem terminus	+		2	1488	0,2731	3,40E+038	1,4486	0,2809
242	volt	+	+	+	+		+		2	3562	0,1141	3,40E+038	1	0,0863
		133	216	190	114									